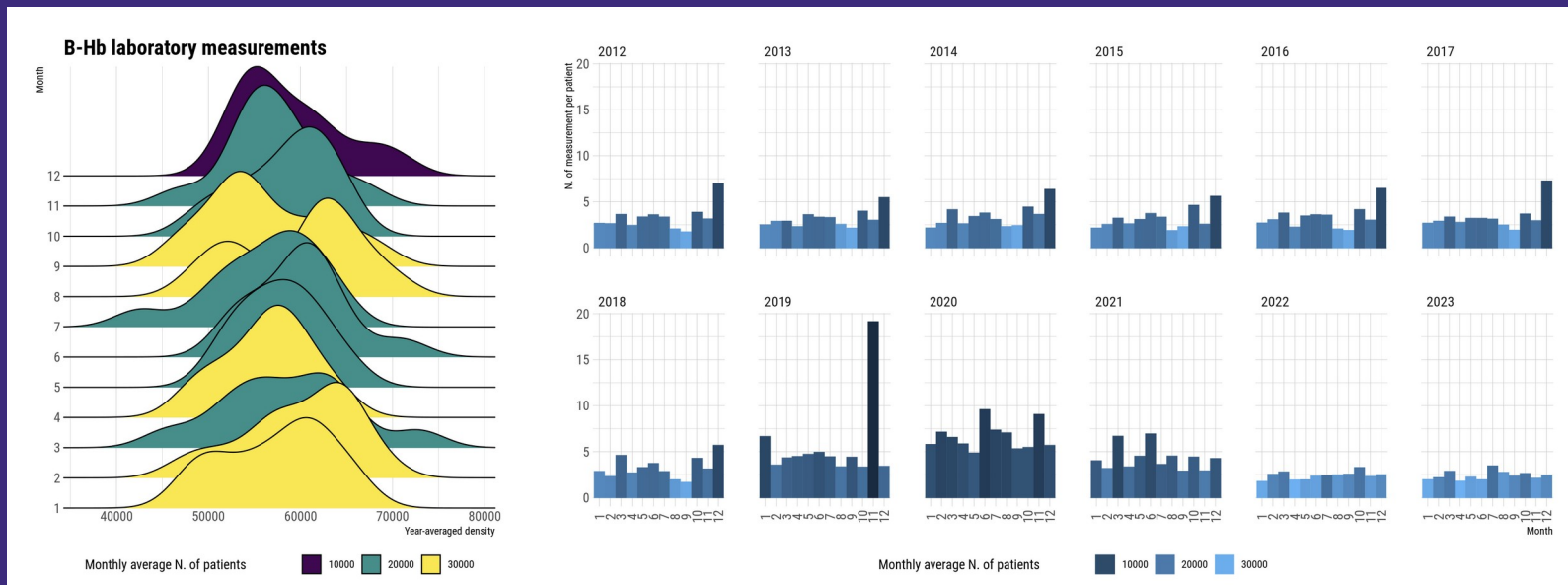




Using Healthcare Data for Scientific Research and Data-Driven Care: Examples from Finland

Arho Virkki

Chief Analytics Officer, Wellbeing services county of Southwest Finland (Varha)

Adjunct Professor in Mathematics, University of Turku

Contents

Background

- Turku University Hospital, Auria Clinical Informatics Unit, and the Wellbeing services county of Southwest Finland (Varha)
- The Finnish way of using health data for the secondary use: History, current practices and future Directions, including EHDS

Electronic Health Records (EHRs) and their Daily Use in Finland

- What does the EHR system in Finland look like?
- What kind of data is collected – and who has access to it?
- How do physicians use the EHR in their day-to-day work?
- Social Security Reform in Finland 2023 and onwards

Secondary Use of Health Data

- How are routine data (billing data, EHR data, etc.) used for analytics?
- The difference between scientific and operational uses of data
- Secure Processing Environments (SPEs)
- Dashboards and benchmarking
- How do physicians or clinics benefit from the insights from data

Efficient and Safe Health Data Use (in speaker's opinion)

- The greatest barriers in data utilization
- Future directions of data utilization in healthcare
- Discussion



Turku University Hospital and Varha

The wellbeing services county of Southwest Finland - Varha

- Responsible for organizing the regional social and health services and rescue operations (since January 1, 2023)
- Covers 27 municipalities

The duties of the wellbeing services counties include:

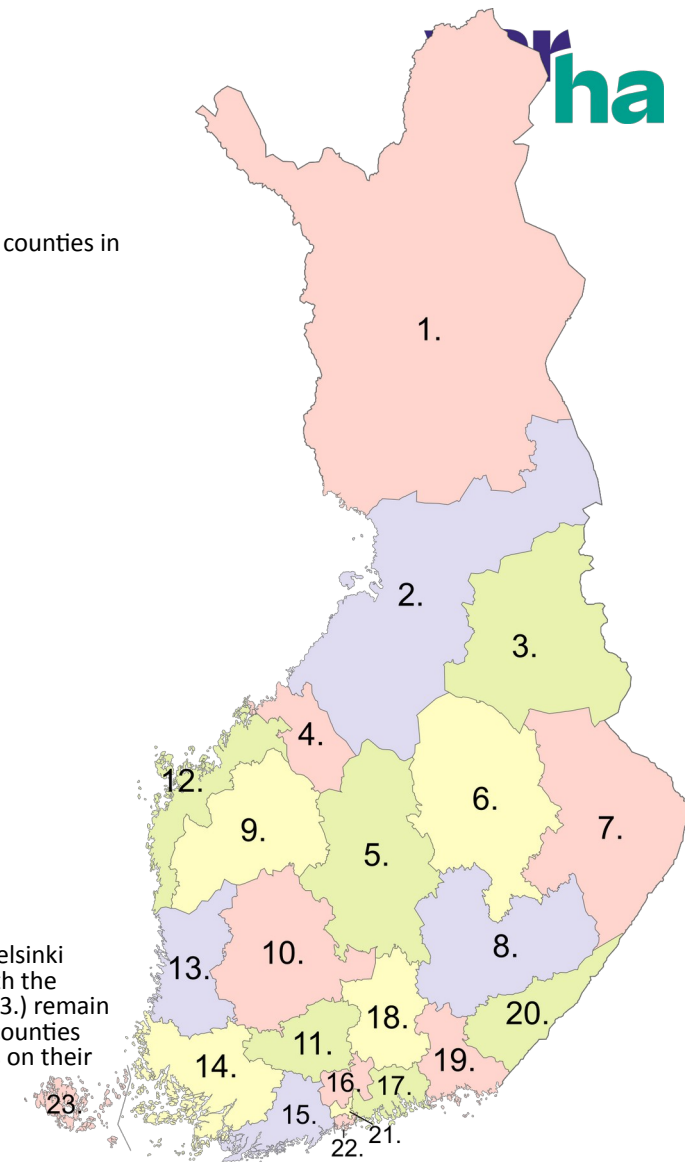
Primary healthcare
Specialised healthcare
Social welfare
Services for children, young people and families
Services for working-age people
Mental health and substance abuse services
Services for persons with disabilities
Student welfare
Rescue services
Prehospital emergency medical services

<https://stm.fi/en/wellbeing-services-counties>

There are 21 wellbeing services counties in Finland:

1. Lapland
2. North Ostrobothnia
3. Kainuu
4. Central Ostrobothnia
5. Central Finland
6. North Savo
7. North Karelia
8. South Savo
9. South Ostrobothnia
10. Pirkanmaa
11. Kanta-Häme
12. Ostrobothnia
13. Satakunta
- 14. Southwest Finland**
15. West Uusimaa
16. Central Uusimaa
17. East Uusimaa
18. Päijät-Häme
19. Kymenlaakso
20. South Karelia
21. Vantaa-Kerava

The City of Helsinki (22.) and Helsinki University Hospital together with the autonomous Region of Åland (23.) remain outside the wellbeing services counties (exercising similar competences on their own.)



Turku University Hospital and Varha

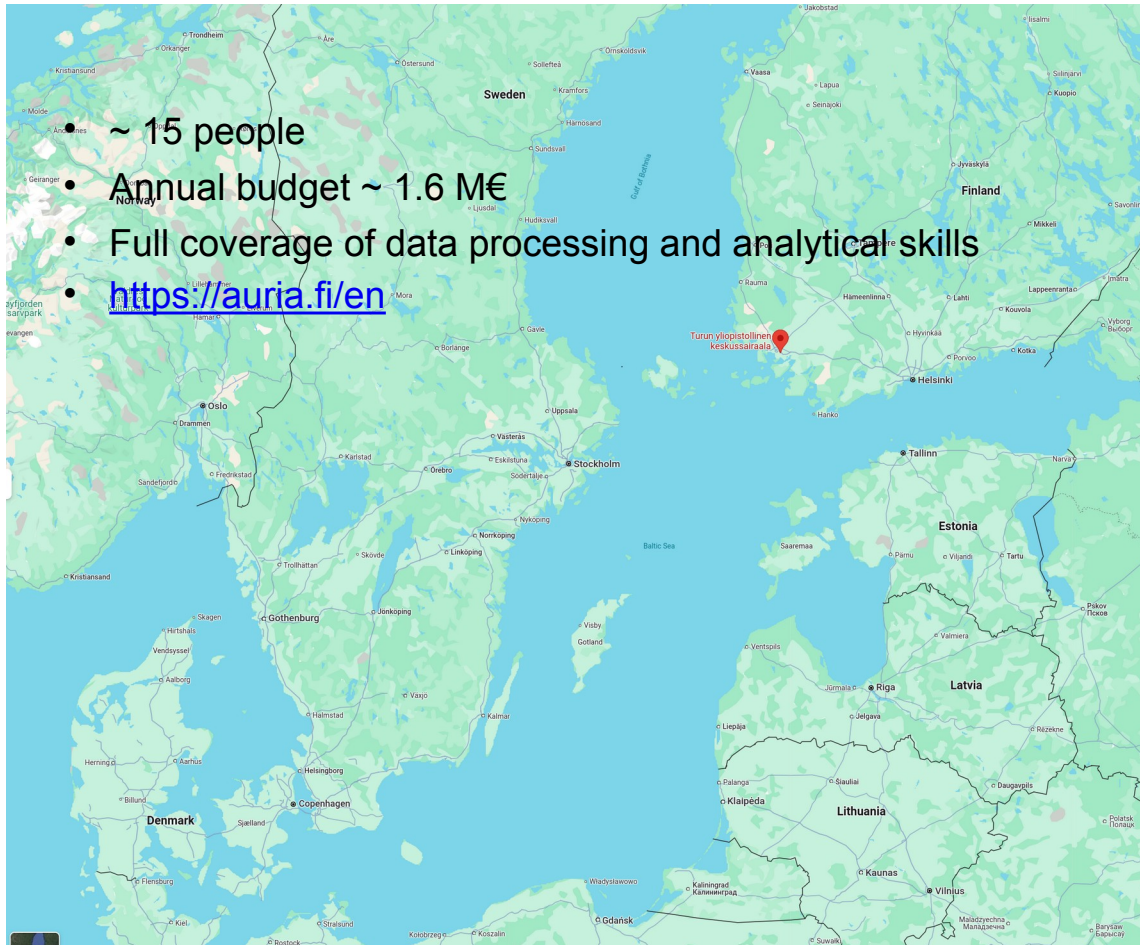
Turku University Hospital

- Was founded in 1756 (and it is the second oldest hospital still in use in the Nordic countries after Rigshospitalet in Copenhagen. The Academy of Turku was founded earlier in 1640)
- Part of Varha since January 1, 2023
- Is used as a teaching hospital by the University of Turku Faculty of Medicine with approx 1500 students in medicine and nursery practice every year.
- Auria Clinical Informatics was started as a project for building a Hospital Data Lake in 2024 and was established as an official service unit in 2019

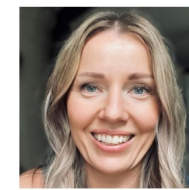


Auria Clinical Informatics

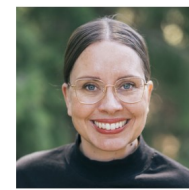
Services for Research and Education



Arho Virkki
Chief Analytics Officer
+358 50 326 2869
arho.virkki@varha.fi



Annika Pirnes
Service Manager
+358 50 336 1925



Lotta Polviander Ekholm
Service Manager
+358 50 329 3893

varha



Antti Yli-Karhu
IT Specialist



Per-Erik Gustafsson
IT Specialist



Tommi Kauko
Senior Statistician



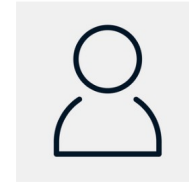
Pia Tajanen-Doumbouya
Database Expert



Harttu Toivonen
Data Architect



Lotta Tapana
Statistician



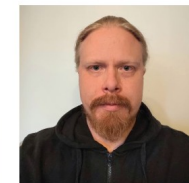
Marjo Kallanmäki
Statistical Programmer



Polytimi Dimitriou
Statistical Programmer



Elina Ahovuo
Statistical Programmer



Matti Leskinen
Data Architect



Altti Tammi
Database Expert

The Finnish way of using health data

1960's

Development of electronic information systems in Finnish health care began in the 1950s and the first health care computer programs were already in use in the 1960s, although then mainly for financial management needs, not for electronic health records (EHRs). Also the Finnish social security id - a unique personal identifier - dates back to 1960's. It has the form DDMMYY-dddC, e.g. 310275-123X or 010103A123Z, where the separator "A" is used for people born in this millennium and "+" for people born in the 19th century.

1970-80's

In the 1970s and 1980s Finland was one of the top countries in the use of information technology internationally, and in hospitals and especially in primary health care. Development of electronic information systems, simultaneously with manual patient record data specification work, has created a foundation for electronic patient records and electronic structured for recording. The first electronic patient record was introduced at the Varkaus Health Centre in 1982.

1990's

To support the development work, the Finnish Hospital Association published a guideline on the structure of a continuous patient record system in 1991, and in the 1990s, electronic patient record systems were comprehensively introduced to Finnish primary health care. At the same time, with the introduction of electronic patient record data, work on the development of the Finnish classification of activities related to electronic structured recording also began.



Image source:
<https://www.veripalvelu.fi/en/history/history-of-the-blood-services-scientific-research-activity/>

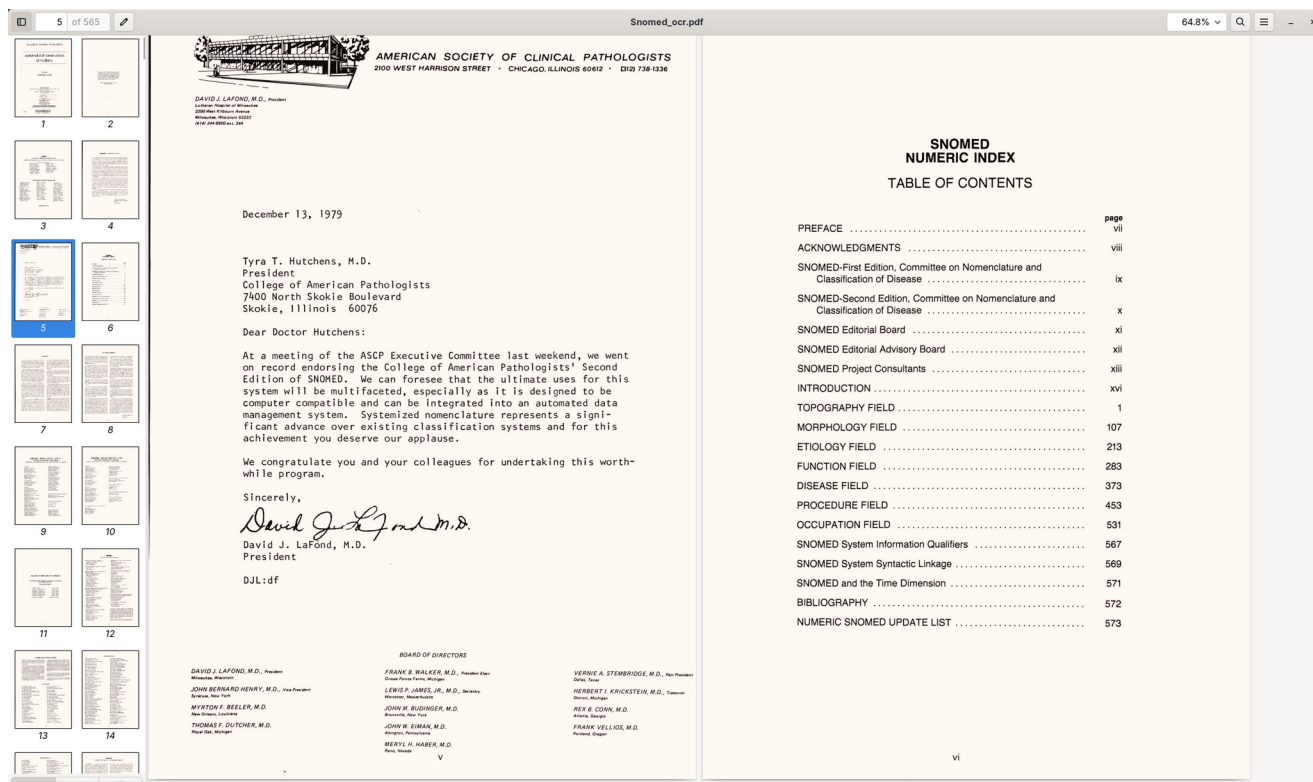


The Finnish way of using health data

For example,

Snomed v2, the systematized nomenclature of medicine from early 1980's is still in use in Finnish Biobanks.

These have been proposals and efforts in converting the historical data into more modern (but not any more hierarchical) Snomed-CT format.



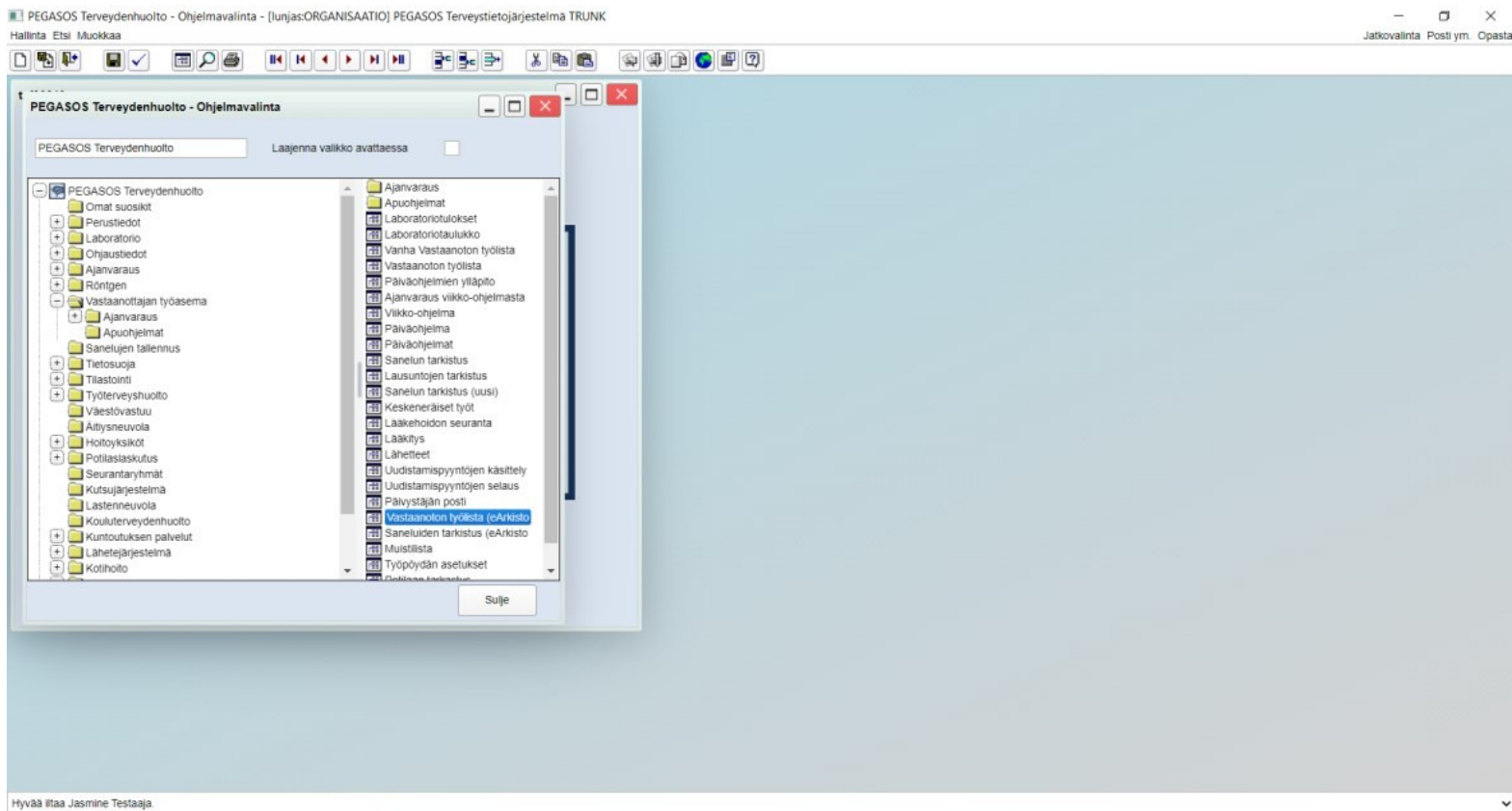
Sources:

- Finnish Biobank Cooperative – FINBB. <https://finbb.fi/en/>
- Carita Olkkonen, The Development of the System and Education of Staff in the Implementation of the Structured Electronic Documentation 29.8.2013 <https://www.theseus.fi/bitstream/handle/10024/64246/Sahkoise.pdf>



The Finnish way of using health data

Semi-random screenshots from the operational EHR systems in use at Southwest Finland

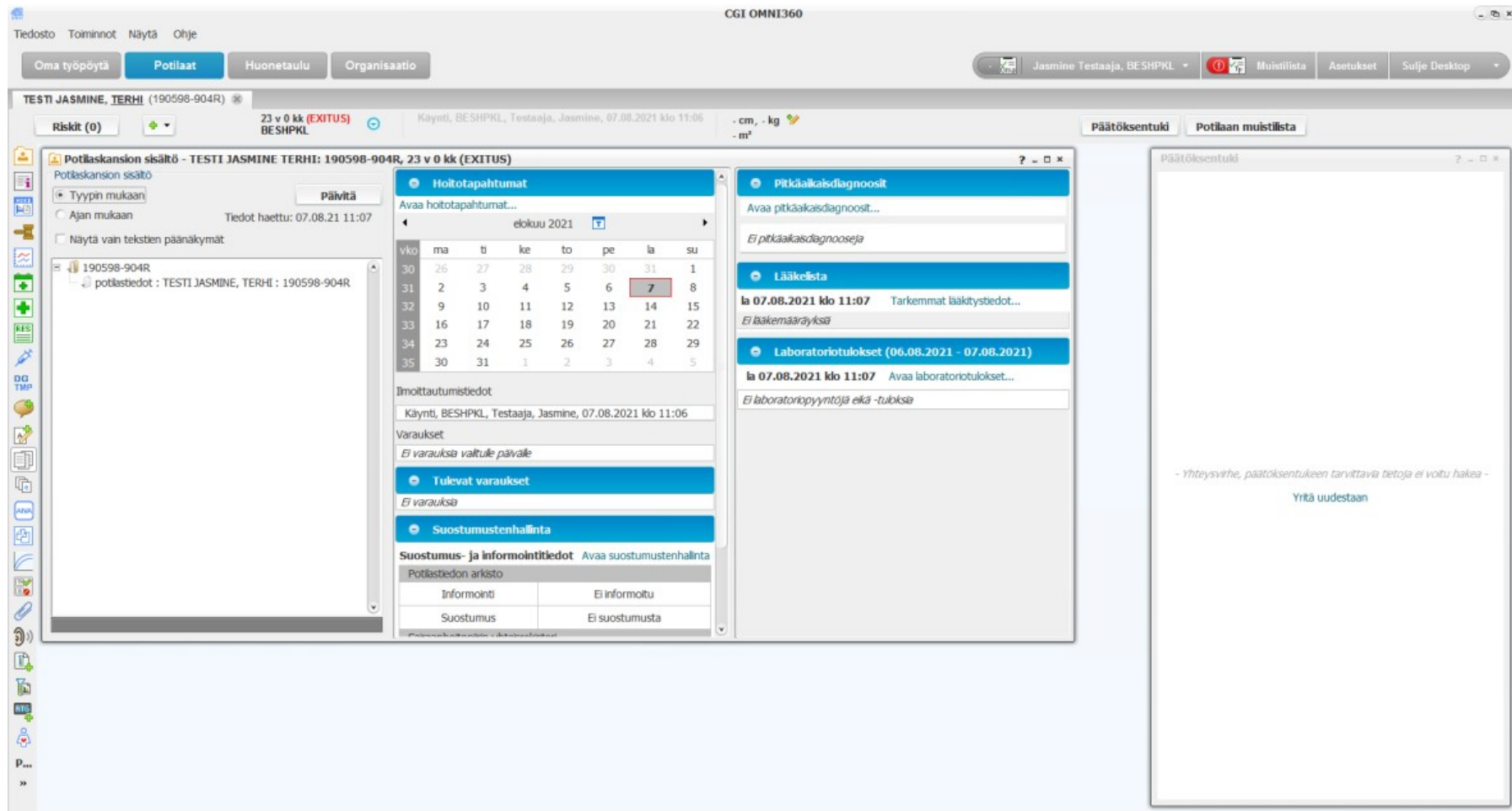


Pegasos system for primary healthcare, main view

Source: Jasmine Lundgren, Bachelor's Thesis, Turku University of Applied Sciences, Business Information Technology, 2021. Usability Differences of Patient Information Systems' Treatment Reports. https://www.theseus.fi/bitstream/handle/10024/507822/Lundgren_Jasmine.pdf

Pegasos system for primary healthcare, notes view

The Finnish way of using health data



The screenshot displays the CGI OMNI360 patient notes view for a patient named TESTI JASMINE, TERHI (190598-904R). The interface is in Finnish and shows various tabs for patient information, including Potilaskansio, Riskit, and Päätöksentuki. The main content area is divided into several sections:

- Potilaskansio sisältö - TESTI JASMINE TERHI: 190598-904R, 23 v 0 kk (EXITUS) BESHPLK**: This section contains a list of patient records and a table of dates.
- Hoitotapahtumat**: A table showing treatment events for the month of August 2021.
- Pitkäaika diagnostiikka**: A section for long-term diagnosis, including a list of diagnoses and a table of results.
- Lääkeliista**: A section for medication, including a list of medications and a table of results.
- Laboratoriotulokset**: A section for laboratory results, including a list of results and a table of results.

The interface also includes a sidebar with various icons for navigation and a top bar with tabs for Oma työpöytä, Potilaat, Huonetaulu, and Organisaatio.

CGI OMNI360
patient notes
view



CGI OMNI360
patient notes
view

The Finnish way of using health data

In 1989, The Act on Special Hospital Care (1062/89) defined three important tasks for Finnish university hospitals:

- 1) special level medical care,
- 2) teaching and
- 3) scientific research.

University hospitals teach doctors, nurses and other health care professionals (e.g. therapists, psychologists). In addition, university medical faculties conduct a lot of small-scale, but medically important, researcher-oriented applied and basic research that utilizes information from electronic health records.

Thus, in addition to patient care, the data from the ERHs is used also

- In connection with advanced studies that are mandatory for medical education, if the student chooses a registry study as the method.
- In a researcher-oriented register study, where, for example, part of the data may be asked based on consent and the extensive control material to be mirrored is obtained based on the secondary law.
- In the work of a doctor, where one's own actions and the effectiveness of treatment decisions are examined in the light of history in order to develop one's own competence.
- When developing the service system as part of the organization's data-driven management.

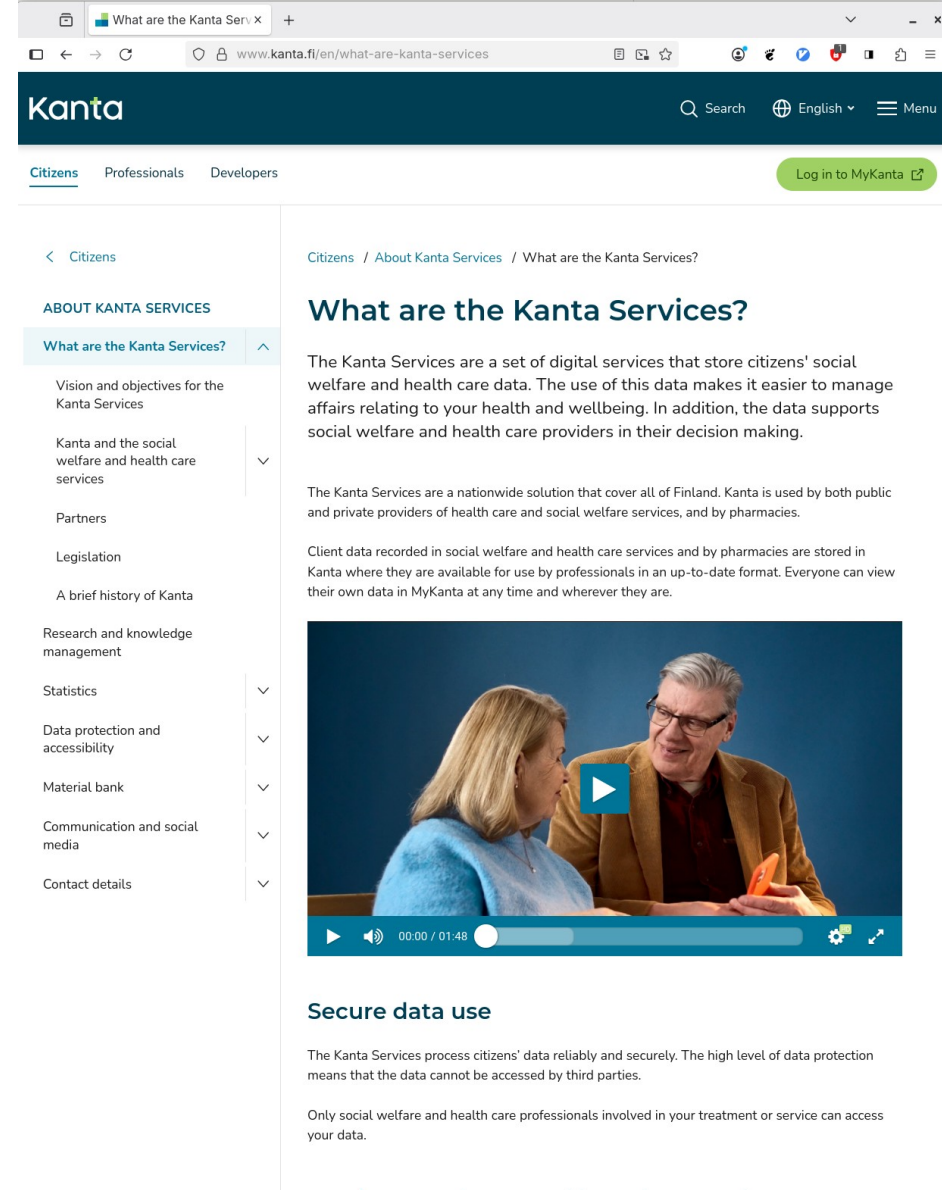
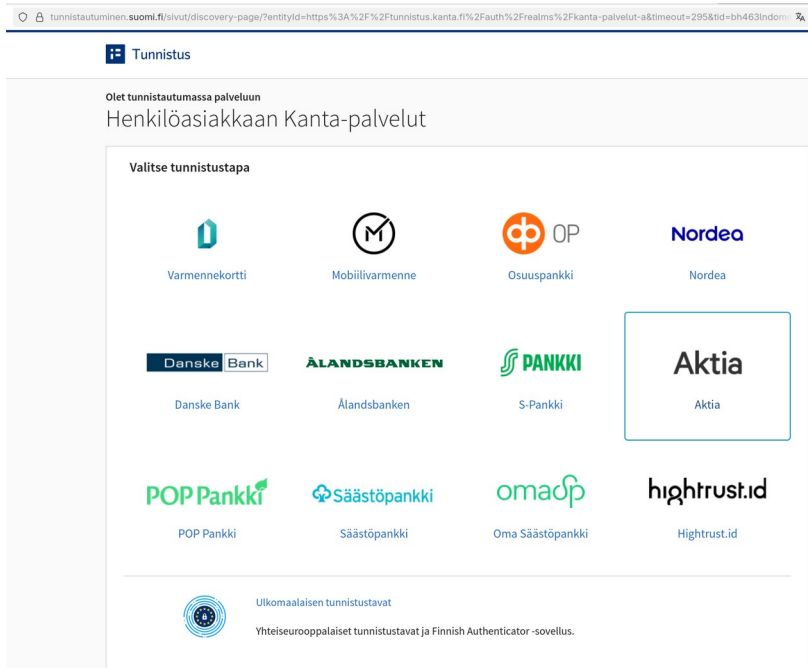


The Finnish way of using health data

2010

The use of electronic prescription via Finnish national Kanta Services became mandatory in 2017.

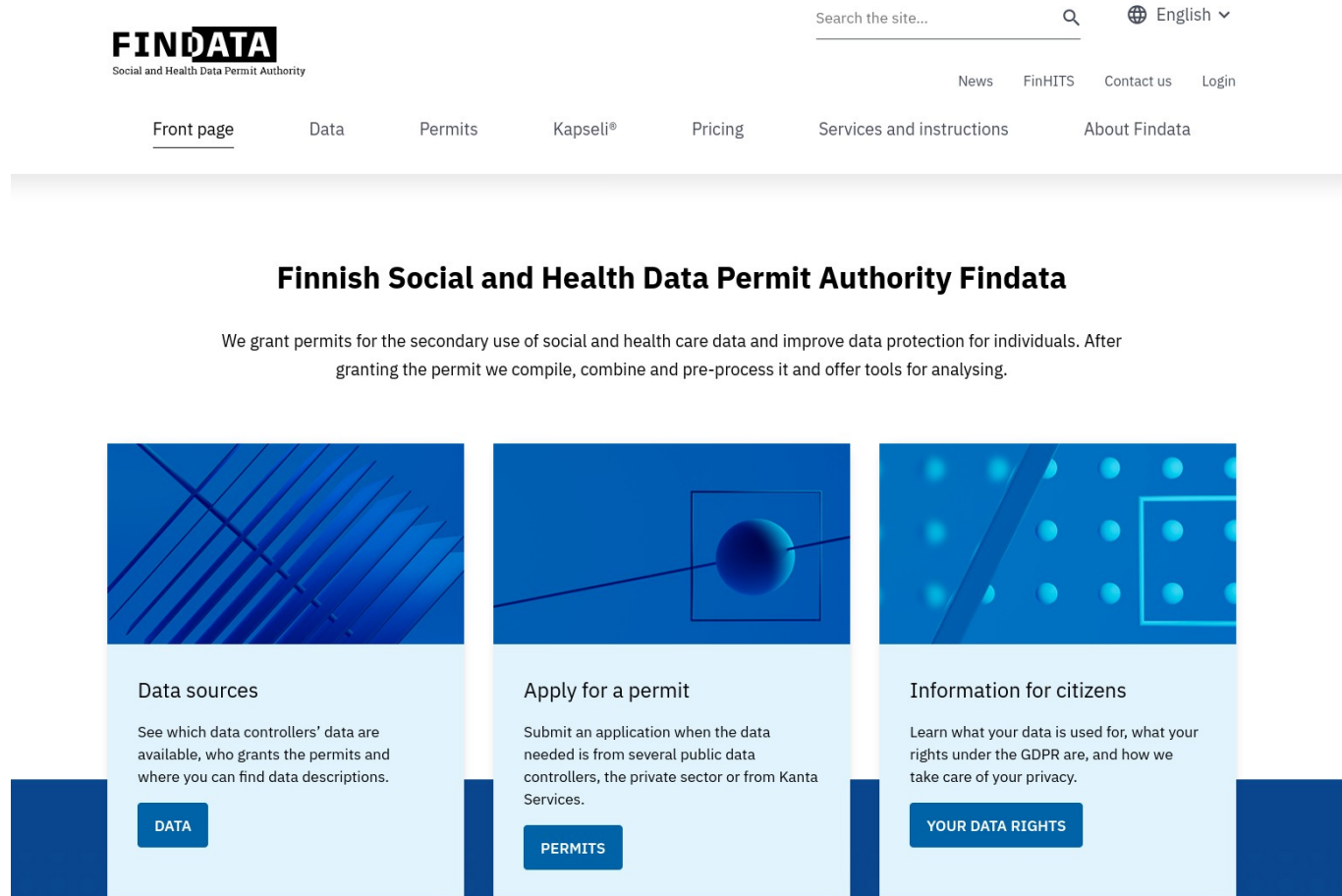
Kanta is hosted by **Kela**, the Social Insurance Institution of Finland, which is supervised directly by the Finnish Parliament. Kela is thus an independent social security institution with its own administration and finances.



The Finnish way of using health data for the secondary purposes

2019 Findata - a national data access body - started its operations in Finland due to the *Act on the Secondary Use of Health and Social Data*.

See: <https://stm.fi/en/secondary-use-of-health-and-social-data>



The screenshot shows the Findata website, which is the national data access body in Finland. The website has a clean, modern design with a blue and white color scheme. The header includes the Findata logo, a search bar, and a language selector set to English. The main navigation menu includes links to Front page, Data, Permits, Kapseli®, Pricing, Services and instructions, and About Findata. The main content area features a large heading "Finnish Social and Health Data Permit Authority Findata" followed by a brief description of the authority's mission. Below this, there are three prominent blue boxes with white text, each representing a key service: "Data sources", "Apply for a permit", and "Information for citizens". Each box includes a brief description and a button to guide the user to the relevant section.

FINDATA
Social and Health Data Permit Authority

Search the site...  English 

News FinHITS Contact us Login

Front page Data Permits Kapseli® Pricing Services and instructions About Findata

Finnish Social and Health Data Permit Authority Findata

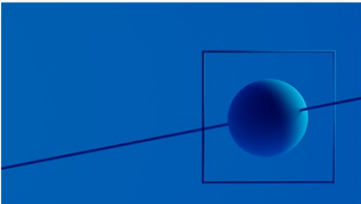
We grant permits for the secondary use of social and health care data and improve data protection for individuals. After granting the permit we compile, combine and pre-process it and offer tools for analysing.



Data sources

See which data controllers' data are available, who grants the permits and where you can find data descriptions.

DATA



Apply for a permit

Submit an application when the data needed is from several public data controllers, the private sector or from Kanta Services.

PERMITS



Information for citizens

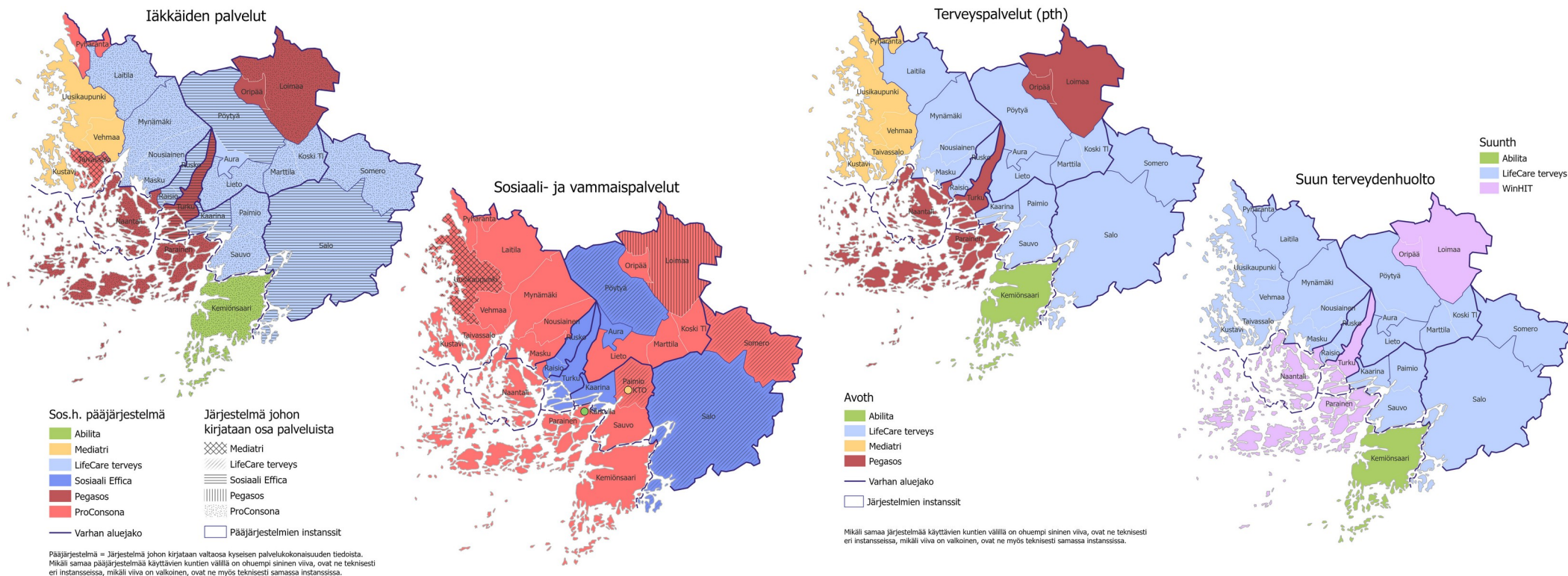
Learn what your data is used for, what your rights under the GDPR are, and how we take care of your privacy.

YOUR DATA RIGHTS

The Finnish way of using health data for the secondary purposes

Current Reality

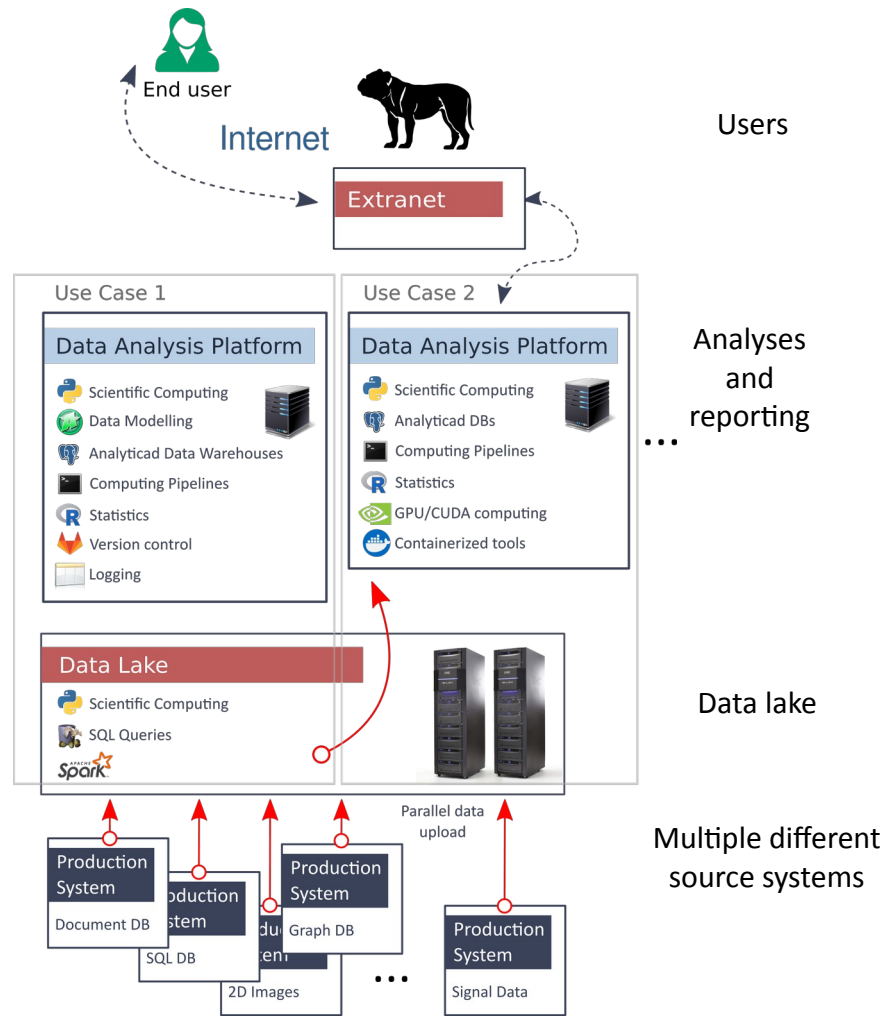
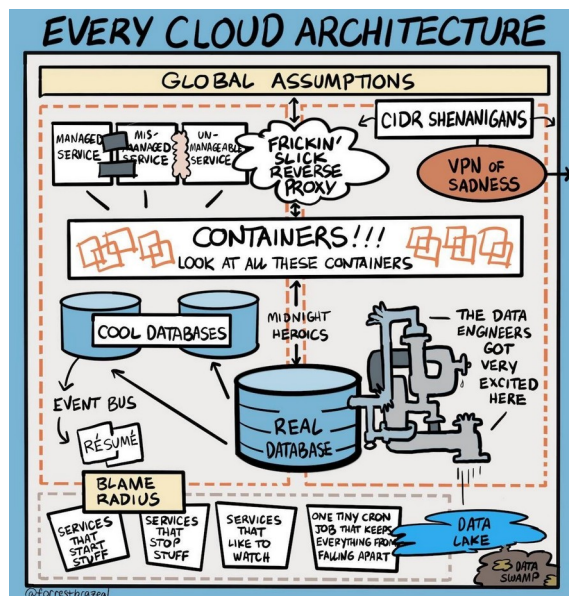
Due to the incremental and distributed history of the ICT systems in health and social services, the amount of different databases and their instances is large. The Social Security Reform, active since the 1st of January 2023 (which was planned since 2005!) aims to consolidate and simplify the current jungle of databases.



The Finnish way of using health data for the secondary purposes

Current practice is to combine data with a layered architecture, which consists of

- 1) Source systems
- 2) Centralized raw data lake
- 3) Different data warehouses (currently, one for reporting and other for scientific data extractions)
- 4) Utilization layers (e.g. reporting or Secure Data Platform, SDP, for science)

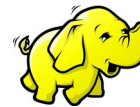
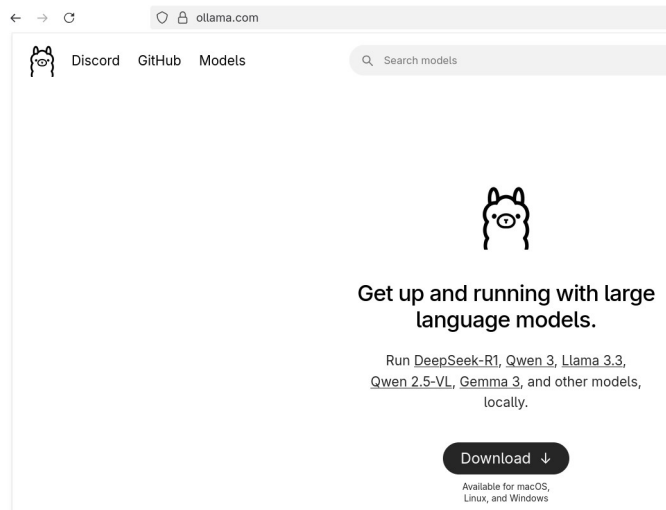


Current Data Infrastructure at Varha

In-house servers with some Azure cloud services for reporting (with aggregated, anonymous data)

Traditional data warehouse technology (SQL Server) for data-driven management and reporting

Free and open sourced (FOSS) solutions for scientific research, development and innovation purposes. => Mixture of traditional and well-established technologies (PostgreSQL and R language for data and Python for orchestration) with varying state-of-art tools, e.g., AI Large Language Models (LLMs) via Ollama and Hugging Face frameworks.



Current Data Infrastructure at Varha

Yes – In-house setup is possible.

At the moment, there seems to be roughly 25 TiB of categorical and text data in our analytical PostgreSQL data warehouse.

The problem of social and healthcare data is not its sheer volume (e.g. it is not that Big Data), but its varying nature and usually free textual form.

```
$ df -h -x tmpfs -x devtmpfs
```

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/mapper/vg_sys-lv_root	207G	36G	171G	18%	/
/dev/sda2	1014M	319M	696M	32%	/boot
/dev/sda1	200M	6.9M	193M	4%	/boot/efi
/dev/sdb1	42T	25T	17T	60%	/data



Auria Clinical Informatics Mission and Vision

Mission:

We support data-driven scientific research in healthcare, which contributes to better and more efficient future treatments and health services.

We provide insight and services for national and EU-level in

- 1) Scientific research,
- 2) Development activities,
- 3) Innovation work and
- 4) Education.

Vision:

- Digitalization and automation - including AI - will change society thoroughly during the next 30 years
- The data should be used for the best of the people. Thus, information security means that confidential health data *is secured to be available for scientific research and medical decision support* in legal and ethical way.
- The only way for Europe to stay relevant is to invest in people and knowledge
- Digital Sovereignty, cost effectiveness and future-proof IT-infrastructure is best ensured, at least in R&D, by using Free and Open Sourced Software (FOSS).

In 2025, we

- Embrace AI – in a sovereign way (open sourced models and local server hall)
- Develop Atolli Secure Processing Environment (SPE) towards AI and Federated Learning tasks
- Focus on external scientific funding for 2026 – 2028 to enable these goals



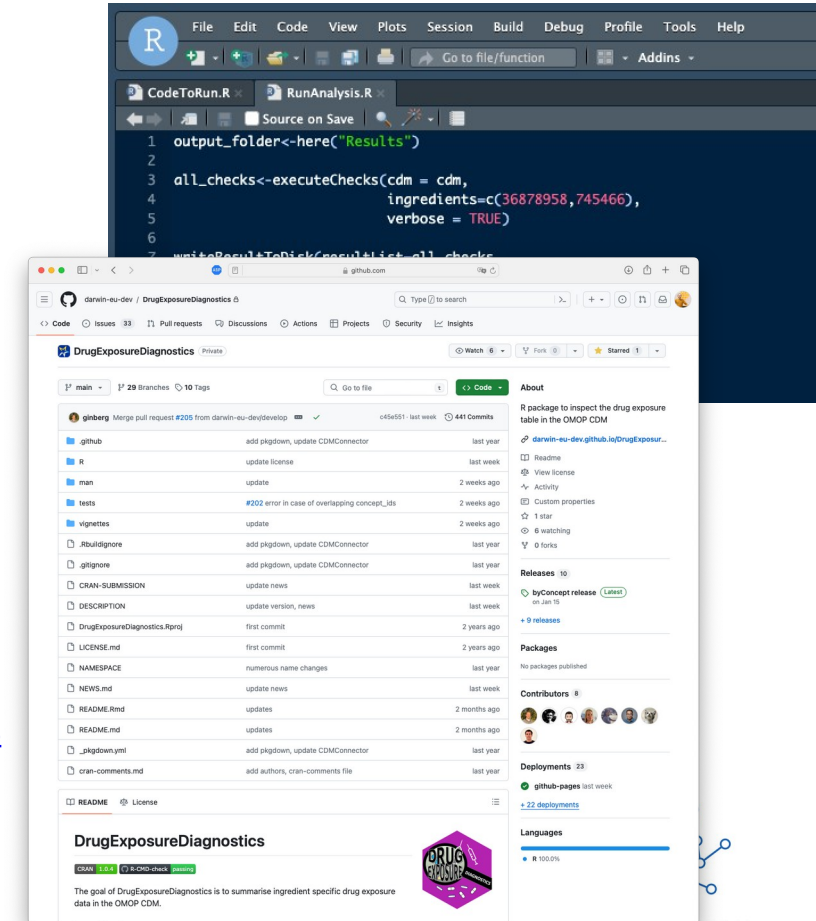
Auria Clinical Informatics Services

Auria Services:

- Data advisory service
- Secure data processing environment (SPE), [Auria's Atolli](#), required by Finnish law
- Data extractions and processing in accordance with data permits granted by Findata and Varha (including cost estimates, advice on data permits)
- Project-funded biostatistical, data processing and scientific computing services
- Biostatistical analyses as contract research
- Data for regulatory clearances

Auria Clients:

- Varha researchers and project personnel, Auria Biobank
- Academic researchers from outside Varha
- Pharmaceutical companies and clinical research organizations (CROs) serving pharmaceutical companies
- Public authorities such as Finnish [Ministry of Social Affairs and Health](#), [Finnish Institute for Health and Welfare](#) and [Findata](#)
- Regulatory authorities such as [Fimea](#) and the [European Medicines Agency EMA](#)
- Consortiums and foundations, such as the EHDEN community formed around the international OMOP data model



In-House Knowledge:

- **Research services:** Understanding the needs of academic and commercial research and strong service expertise, which is reflected in the positive images associated with the customer and the Auria brand during the rapid response times
- **Data extractions:** High-quality formation of research material in accordance with the service request and research plan, as well as a broad understanding of the content and technical structure of the material. Excellent technical ability to develop knowledge-picking methods. Close and good contact with clinical experts at Varha and other registrars.
- **Data management:** Data technical chapter, aggregation, commensurability and code conversions (such as OMOP format) and quality assurance methods
- **Data analytics:** KPIs, methods of statistical reasoning and prediction (e.g., biostatistics of drug efficacy and safety evaluation) and nonparametric statistics, e.g., artificial intelligence applications (e.g., for the needs of research and development projects).
- **Scientific computing:** Audited secure processing environment Technologies required by atolls, such as applications of computer networks, servers, software and protocols, and cryptography. Technical security and high-performance computing applications, including GPU computing required by artificial intelligence.
- **Data protection:** In addition to technical and organizational security, statistical disclosure control methods such as knowledge of differential data protection and anonymization capabilities for different data. In-depth knowledge of Finnish data protection legislation and official position in its interpretation as part of data protection groups in secondary schools

blackbird.vssh.net (CentOS Linux 7.9.2009 64bit / Linux 3.10.0-1160.81.1.el7.x86_64)

Uptime: 638 days, 15:41:28

```

Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz - 1.94/3.70GHz
CPU [ ] 4.6% idle 95.5% ctx_sw 41K
MEM [ ] 18.3% active 8736
SWAP [ ] 0.0% 0.0% inter 5746
total 1.47T inacti 5746
used 276G buffer 976K
free 1.21T cached 1.20T
LOAD - 96core
1 min 4.14
5 min 8.61
15 min 3.64

```

```

NETWORK Rx/s Tx/s TASKS1056 (3007 thr), 5 run, 1051 slp, 0 oth Threads sorted automatically by CPU consumption
br912 1Kb 2Kb
br-34ef7df938cd 0b 0b
br-a68c42e98757 0b 0b
doker0 1Kb 1Kb
eth3 3Kb 4Kb
lo 376b 376b
team0 3Kb 4Kb
team0_912 2Kb 4Kb
veth1b5417a 0b 0b
veth2be9909 0b 0b
veth2c64275 0b 0b
veth2fccfa0 0b 0b
veth5c3a688 0b 0b
2024-10-16 12:06:54 EEST 0b 0b

```

Data Types

1. Numeric and categorical data and free text such as

Demographic variables, gender, date of birth, time of death, municipality of residence and address, hospital district, mother tongue, profession, days spent in the hospital (department, specialty, reason for admission, arrival and departure date), from and where the patient was transferred, IDC10-coded diagnoses, surgery information (anesthetics and doses), pathological diagnoses (topology and morphology), radiotherapy data, imaging studies (CT images, magnetic resonance images, PET images,...), neurophysiological studies, prescriptions and drugs administered in the hospital (brand name, active substance and dose), laboratory studies, nurses' and doctors' records, hospital infections, spirometry tests,...

AZ resepti_numero	AZ atc_koodi	AZ atc_selite	AZ laakeaine	AZ laakeainetyyppi	AZ kauppanimi	aika_pvm	AZ annostelu
569855160689991157	J05AB14	Valgansikloviiri	valgansikloviirihydrokloridi	vaikuttava_aine	VALCYTE ORIFARM	2015-09-12	1 tabl/vrk, verenkiertolääke
305542956377470999	J01FF01	Klindamysiini	klindamysiinihydrokloridi	vaikuttava_aine	DALACIN	2013-09-16	1 tabletti aamulla ja 2 tablettia illalla
665843149687446861	P01BA02	[NULL]	hydroksiklorokiini	vaikuttava_aine	OXIKLORIN	2017-05-09	D.S. 3 + 3 tabl/vrk suolitulehduksen hoitoon
252954761730757998	S01AE05	Levofloksasiini	levofloksasiinihemihydraatti	vaikuttava_aine	OFTAQUIX	2010-09-30	1 tabletti 2 kertaa päivässä
320820191458508974	N06AX21	Duloksetiini	duloksetiinihydrokloridi	vaikuttava_aine	DULOXETINE ORION	2015-11-06	1-2 droppar x 6 / dag i vänster öga börjanc
735543039365900410	C03DA01	Spironolaktoni	spironolaktoni	vaikuttava_aine	SPIRONOLACTONE	2011-07-21	2 tbl x 3
098907713971128485	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	2015-03-05	perusvoide, pitkäaikaisen ihosairauden hoit
921156162038303037	A02BC02	Pantopratsoli	pantopratsoli	vaikuttava_aine	SOMAC	2013-10-25	1 tbl aamuisin. Verenpainelääke.
901303991418620388	[NULL]	[NULL]	CERIDAL SUIHKE	kauppanimi	CERIDAL SUIHKE	2012-01-30	1-2 tabl. päivässä.
506054736480710795	N02AX02	[NULL]	tramadolihydrokloridi	vaikuttava_aine	TRAMAL RETARD	2011-03-13	1 tippa aamuisin oikeaan silmään
576889490718991216	N05BA02	Klooridiatsepsoksiidi	klooridiatsepsoksiidi	vaikuttava_aine	RISOLID	2013-06-14	1 tabl 1-3 krt tarv. kipuun
437380735261683006	N02AX02	Tramadoli	tramadolihydrokloridi	vaikuttava_aine	TRAMAL RETARD	2015-01-01	10 mg x 1 PO
082413491707888845	L01BA01	Metotreksaatti	metotreksaatti	vaikuttava_aine	TREXAN	2015-05-13	1 tabletti kerran päivässä. Rintasyövän horn
440604585409710408	J01DB01	[NULL]	kefaleksiini	vaikuttava_aine	KEFEXIN	2016-10-02	35 yksikköä aamulla, 20 yksikköä päivällisell
177109125210616727	A12AX	Kalsiumin yhdistelmävalmisteet D-vitamiinin ja/tai m	kalsiumkarbonaatti,kolekalsiferoli (pulvis)	vaikuttava_aine	KALCIPOS-D FORTE	2012-11-28	Laskevat annokset: 40 mg x1 1 viikko, jonka
879964130395236080	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	2014-03-06	1x1
930714871742065621	A06AD15	Makrogoli	makrogoli 4000	vaikuttava_aine	PEGORION	2014-10-28	1x3, kuuri loppuun
437212213616936136	M03BX02	Titsanidiini	titsanidiinihydrokloridi	vaikuttava_aine	SIRDALUD	2014-02-05	1 tabletti kolmesti päivässä kivun hoitoon.
567253713653239173	S02AA03	Boorihappo	boorihappo,etanoli (96 %)	vaikuttava_aine	OTIBORIN FORTE	2014-05-22	1 tabletti 3 kertaa päivässä sytostaattien aik
536453260582631684	D11AH01	Takrolimuusi	takrolimuusi	vaikuttava_aine	PROTOPIC	2016-09-21	Yksi tabletti kahdeksan tai kahdentoista tur

Data Types

2. Signal data such as

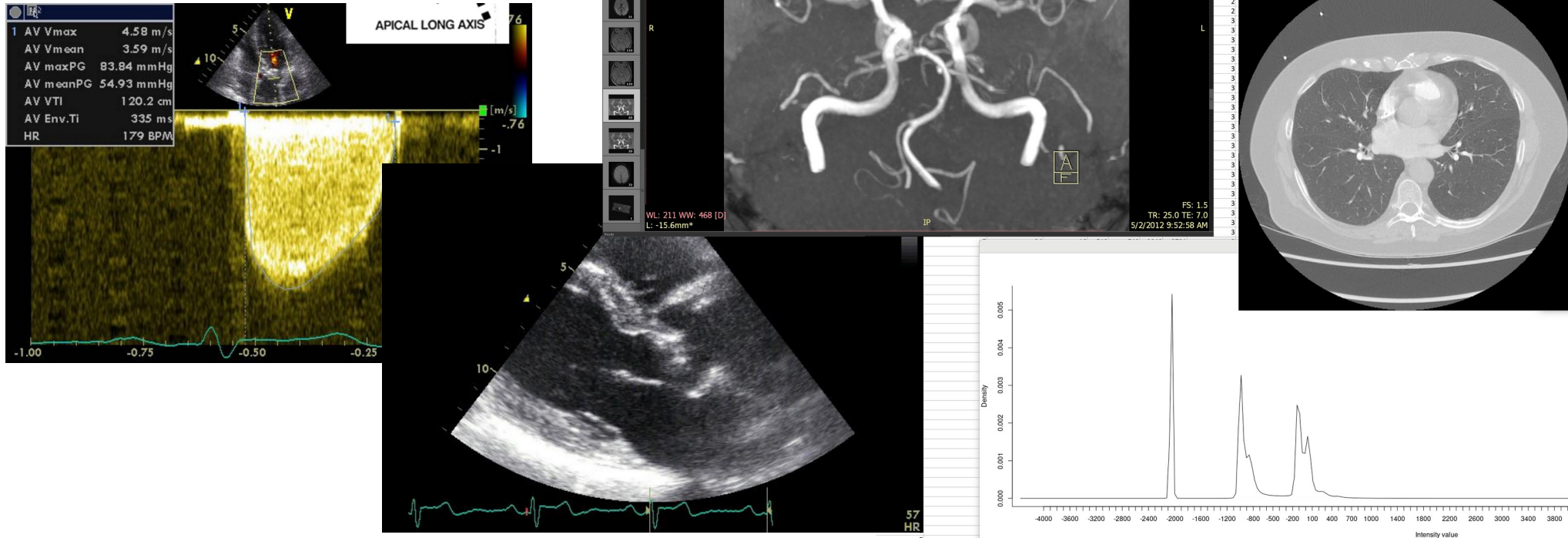
Intensive care unit biosignals (pulse, electrocardiogram, pain-indicating muscle tension signals,...), remote monitoring signals (pressure curve produced by the CPAP device and periods of apnea)



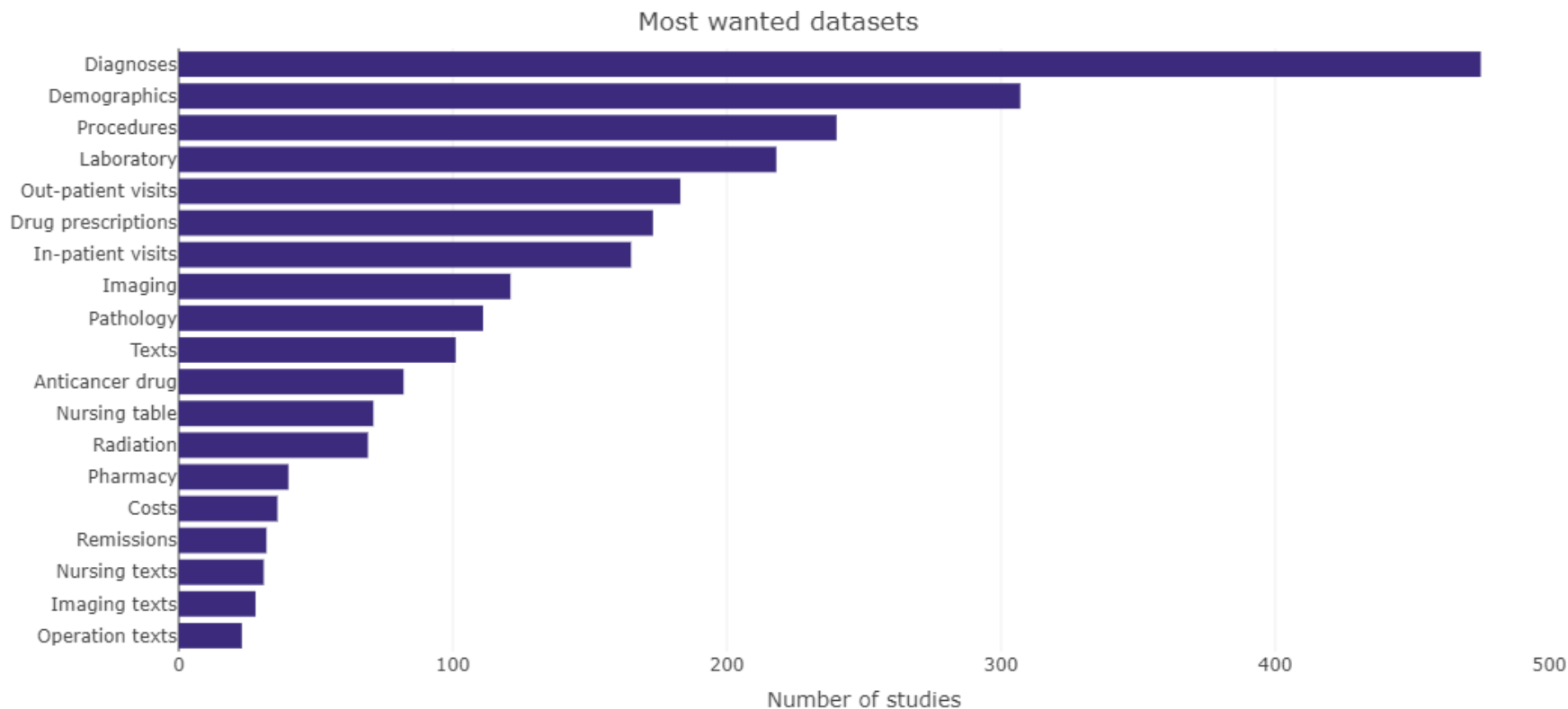
Data Types

3. 2D- and 3D-images

Digital pathology, ultrasound imaging, CT images, magnetic resonance imaging, PET images,...



The Most Frequently Used Data Types in Research



Uses of Patient Data:

Typical Pharmaceutical Company-Oriented Studies

Need for real-life data	Research outcome	Beneficiaries
Understand the burden of disease and current treatment practices	Unmet medical need, the size of target population	Public authorities*, taxpayers, clinicians
Understand current healthcare costs	A health economic model of treatment effectiveness	Health technology assessment (HTA), payers
Assemble a control cohort for a clinical drug trial	Efficacy and safety information for the marketing authorization application	Marketing authorization authority
Show product efficacy and safety vs. clinical trials and alternative treatments	Real-world efficacy and safety information	Marketing authorization authority, clinicians
Show the actual cost effectiveness of the product	Evidence for the price and reimbursement application (or renewal).	Public authority
Shows the preliminary applicability of the product to other than the original target population/indication	Real-world efficacy and safety information outside the original target group	Public authority, payers Marketing authorization authority, Clinicians



Secure Processing Environment (SPE) Auria Atolli



Features

- Linux operating system
- Browser-based remote desktop (independent of client machine, no plugins or client software needed)
- Easily customizable to most needs, often installed additional programs SPSS and SAS
- Atoll website (standard software, machine packages, price lists): <https://www.auria.fi/tietopalvelu/en/atolli/index.html>

User Statistics:

- Environments ordered in total 51
- Total number of users about 200

Terms and Conditions:

- A data permit under secondary law must be valid
- The service is paid

Contact persons:

- Research material advice: tutkimuksentietopalvelut@varha.fi
- Product owner: Service Manager Annika Pirnes, annika.pirnes@varha.fi
- Responsible officer: Arho Virkki, Director of Analytics, arho.virkki@varha.fi



AURIA CLINICAL
INFORMATICS

Atolli's Audit

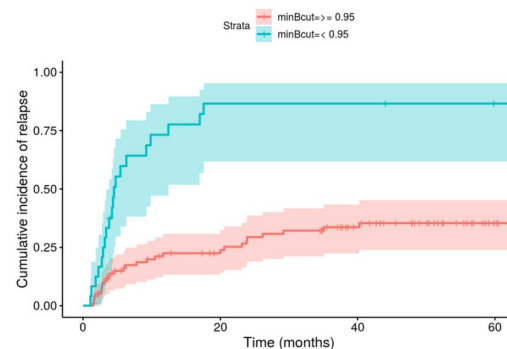
Some pictures taken on 2022-09-28 at the Turku University Hospital during Atolli service audit by KPMG



Atolli supports both traditional and modern statistics

Relapse	no (N=73)	yes (N=49)	Overall (N=122)
Gender			
female	28.0 (38.4%)	25.0 (51.0%)	53.0 (43.4%)
male	45.0 (61.5%)	24.0 (49.0%)	69.0 (56.6%)
Age at transplantation (years)			
Median (Q1, Q3)	53.0 (42.0, 60.0)	57.0 (46.0, 62.0)	55.0 (43.0, 61.0)
Min, Max	19.0, 69.0	23.0, 68.0	19.0, 69.0
Diagnosis class			
g1	30.0 (41.1%)	29.0 (59.2%)	59.0 (48.4%)
g2	7.00 (9.6%)	5.00 (10.2%)	12.0 (9.8%)
g3	5.00 (6.8%)	3.00 (6.1%)	8.00 (6.6%)
g4	31.0 (42.5%)	12.0 (24.5%)	43.0 (35.2%)
Donor			
mud	59.0 (80.8%)	36.0 (73.5%)	95.0 (77.9%)
sibling	14.0 (19.2%)	13.0 (26.5%)	27.0 (22.1%)
Graft source			
blood	64.0 (87.7%)	46.0 (93.9%)	110 (90.2%)
bone marrow	9.00 (12.3%)	3.00 (6.1%)	12.0 (9.8%)
Graft cell count			
Median (Q1, Q3)	6.00 (4.29, 9.20)	6.48 (4.70, 8.76)	6.34 (4.70, 9.09)
Min, Max	0.750, 15.6	1.00, 13.5	0.750, 15.6
Missing	2.00 (2.7%)	0 (0%)	2.00 (1.6%)
Condition regimen			
MAC	26.0 (35.5%)	14.0 (28.6%)	40.0 (32.8%)
RIC	34.0 (46.5%)	22.0 (44.9%)	56.0 (45.9%)
sequential	13.0 (17.8%)	13.0 (26.5%)	26.0 (21.3%)
DRI stage			
high	21.0 (28.8%)	18.0 (36.7%)	39.0 (32.0%)
low	52.0 (71.2%)	31.0 (63.3%)	83.0 (68.0%)
EBMT risk score			
Median (Q1, Q3)	4.00 (3.00, 5.00)	3.00 (3.00, 5.00)	4.00 (3.00, 5.00)
Min, Max	1.00, 6.00	1.00, 6.00	1.00, 6.00

Minimum B Ball



```

Median survival with 95% CI
median lower upper
[1,] "minBcut>= 0.95" "NA" "NA" "NA"
[2,] "minBcut< 0.95" "4.50" "3.80" "12.42"

Log-rank test

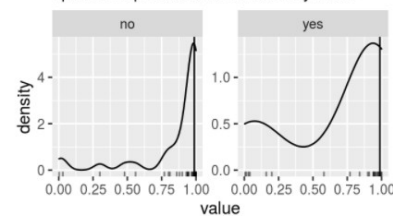
Call:
survdiff(formula = Surv(relapstime, relapse_stat) ~ minBcut,
data = analysisdata)

N Observed Expected (O-E)^2/E (O-E)^2/V
minBcut>= 0.95 99 29 42.4821 4.2787 32.913
minBcut< 0.95 24 20 6.5179 27.8871 32.913

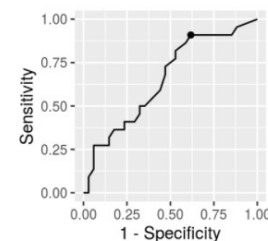
Chisq= 32.9 on 1 degrees of freedom, p= 9.64e-09

```

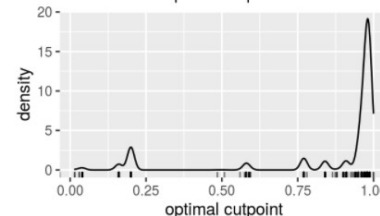
Independent variable
optimal cutpoint and distribution by class



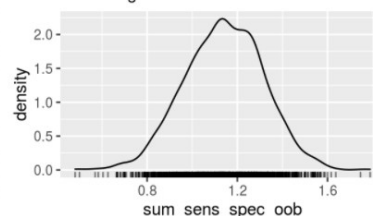
ROC curve



Bootstrap
distribution of optimal cutpoints



Bootstrap
out-of-bag estimates



Some Ongoing Projects



EU Project number: 101095384

Project name: Privacy Compliant Health Data

As A Service For AI Development

Call: HORIZON-HLTH-2022-IND-13



EU Project number: 101083544

Project name: HealthHub Finland - the future of Healthcare shaped by a Hub of partners facilitating data-driven digital solutions in Finland and Europe

Project acronym: HHFIN

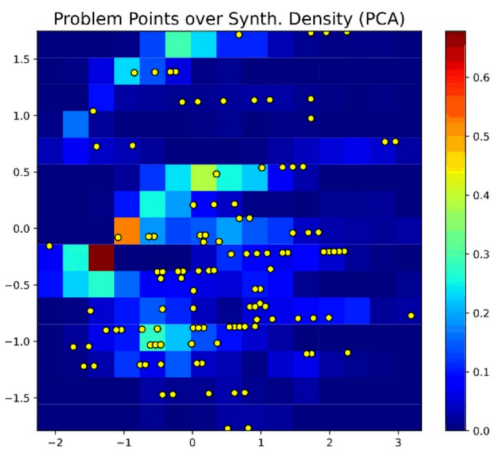
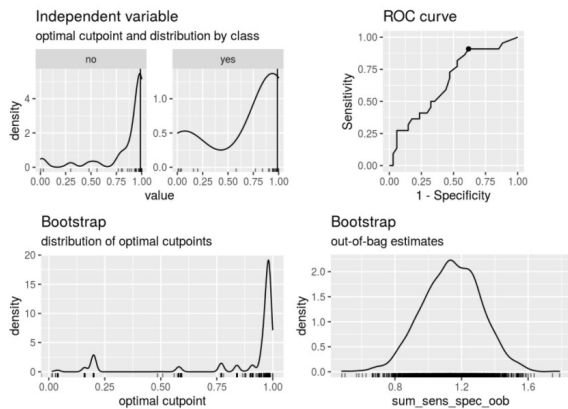
Call: DIGITAL-2021-EDIH-01

Topic: DIGITAL-2021-EDIH-INITIAL-01

The test platform for synthetic health data supports new kinds of research, development, and innovation activities. The project is carried out by the University of Turku, Turku University of Applied Sciences, Southwest Finland Welfare Area and Business Turku.

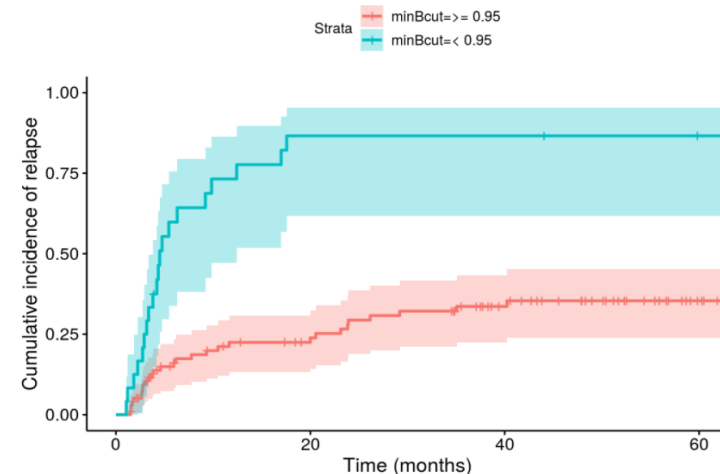
The project is co-funded by the European Union. The project is funded by the Regional Council of Southwest Finland under the European Regional Development Fund (ERDF) Programme for Regional and Structural Policy 20212027. The authority supervising the funding is the Uusimaa Regional Council

Examples of Research



Relapse	no (N=73)	yes (N=49)	Overall (N=122)
Gender			
female	28.0 (38.4%)	25.0 (51.0%)	53.0 (43.4%)
male	45.0 (61.6%)	24.0 (49.0%)	69.0 (56.6%)
Age at transplantation (years)			
Median (Q1, Q3)	53.0 (42.0, 60.0)	57.0 (46.0, 62.0)	55.0 (43.0, 61.0)
Min, Max	19.0, 69.0	23.0, 68.0	19.0, 69.0
Diagnosis class			
g1	30.0 (41.1%)	29.0 (59.2%)	59.0 (48.4%)
g2	7.00 (9.6%)	5.00 (10.2%)	12.0 (9.8%)
g3	5.00 (6.8%)	3.00 (6.1%)	8.00 (6.6%)
g4	31.0 (42.5%)	12.0 (24.5%)	43.0 (35.2%)
Donor			
mud	59.0 (80.8%)	36.0 (73.5%)	95.0 (77.9%)
sibling	14.0 (19.2%)	13.0 (26.5%)	27.0 (22.1%)
Graft source			
blood	64.0 (87.7%)	46.0 (93.9%)	110 (90.2%)
bone marrow	9.00 (12.3%)	3.00 (6.1%)	12.0 (9.8%)
Graft cell count			
Median (Q1, Q3)	6.00 (4.29, 9.20)	6.48 (4.70, 8.76)	6.34 (4.70, 9.09)
Min, Max	0.750, 15.6	1.00, 13.5	0.750, 15.6
Missing	2.00 (2.7%)	0 (0%)	2.00 (1.6%)
Condition regimen			
MAC	26.0 (35.8%)	14.0 (28.6%)	40.0 (32.8%)
RIC	34.0 (46.8%)	22.0 (44.9%)	56.0 (45.9%)
sequential	13.0 (17.8%)	13.0 (26.5%)	26.0 (21.3%)
DRI stage			
high	21.0 (28.8%)	18.0 (36.7%)	39.0 (32.0%)
low	52.0 (71.2%)	31.0 (63.3%)	83.0 (68.0%)
EBMT risk score			
Median (Q1, Q3)	4.00 (3.00, 5.00)	3.00 (3.00, 5.00)	4.00 (3.00, 5.00)
Min, Max	1.00, 6.00	1.00, 6.00	1.00, 6.00

Minimum B all



Median survival with 95% CI

```
median lower upper
[1,] "minBcut >= 0.95" "NA" "NA" "NA"
[2,] "minBcut < 0.95" "4.56" "3.80" "12.42"
```

Log-rank test

```
Call:
survdifff(formula = Surv(relapsetime, relapse_stat) ~ minBcut,
data = analysisdata)
```

```
N Observed Expected (0-E)^2/E (0-E)^2/V
minBcut >= 0.95 99 29 42.4821 4.2787 32.913
minBcut < 0.95 24 20 6.5179 27.8871 32.913
```

Chisq= 32.9 on 1 degrees of freedom, p= 9.64e-09

SHARED is an active research project exploring opportunities, developing methods, and raising awareness of synthetic data in healthcare.

[Learn More](#)

The untapped potential of synthetic data in health research is a gold mine just waiting to be uncovered.

 **Health data should be used to develop innovative health technology**
Health data is a gold mine of vital knowledge with a unique potential to improve collective health and develop world-class health tech solutions.

 To protect patient privacy, we must respect GDPR

At present, several legal, regulatory, and organizational barriers are restricting the access to health data in Denmark and in the other Nordic countries

 Synthetic data is safe data

The benefit of synthetic data is that, unlike de-identified datasets, synthetic datasets contain no data from natural persons, which completely eliminates the risk of re-identification. But to do that, we need robust methods for synthesizing data. This is what the SHARED project is about.

Olkoon ε ja δ valittu. Mikäli kaikilla $x, y \in \mathcal{D}$, $\|x - y\| \leq 1$ ja $\mathcal{S} \in \text{Range}(\mathcal{M})$ pätee, että

$$\underbrace{P(\mathcal{M}(x) \in \mathcal{S})}_{\text{vastausten } \mathcal{S} \text{ osuus kannasta } x} \leq e^\varepsilon \underbrace{P(\mathcal{M}(y) \in \mathcal{S})}_{\text{vastausten } \mathcal{S} \text{ osuus kannasta } y} + \delta$$

niin silloin $\mathcal{M}$ on (ε, δ) -differentiaalisesti yksityinen. Jos $\delta = 0$, niin $\mathcal{M}$ on ε -differentiaalisesti yksityinen.

Määritelmä on helpompi ymmärtää asettamalla lievennysparametri $\delta = 0$ ja kirjoittamalla epäyhtälö prosessin $\mathcal{M}$ tuottamien vastausten suhteen rajoituksena silloin, jos tietokannat eroavat vain yhden askeleen verran: Mielivaltaiselle vastausten osajoukolle $\mathcal{S}$ pätee, että todennäköisyyksien

suhteellinen ero $\frac{P(\mathcal{M}(x) \in \mathcal{S})}{P(\mathcal{M}(y) \in \mathcal{S})} \leq e^\varepsilon = \text{kiinteä raja, esim. } e^\varepsilon = \frac{3}{2}.$

```
--> <syndate> dp-labra-syndate.sql x

-- In the following, we apply the method presented in the publication - "Royce J Wilson,
-- celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant
-- Gipson. Differentially Private SQL with Bounded User Contribution
-- https://arxiv.org/abs/1909.01917 ja sen implementaatiota, Googlen DP-kirjasto
-- https://github.com/google/differential-privacy. Asennuksesta Aurian tietovarastoon kts.
-- https://gitlab.vsshp.net/aci/wiki-pages/-/wikis/08a_software/google-differential-privacy

-- The idea is simply to introduce an additional cut-off parameter M to limit the maximal
-- contribution of a single person to the aggregate table. For example, if a patient has
-- more than M B-Hb-measurements in her records, we include randomly only M of them. The
-- randomization is implemented by numbering the partition rows in random order, and
-- choosing only the rows 1...M.

-- We create a temp table 'sample' to speed up test computations

drop table if exists sample;
create temp table sample as

--[]

drop table if exists p_sample;
create temp table p_sample as select henkilotunnus, tutkimus_nimike, tulos_teksti,[]

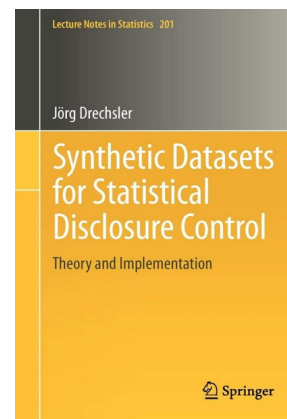
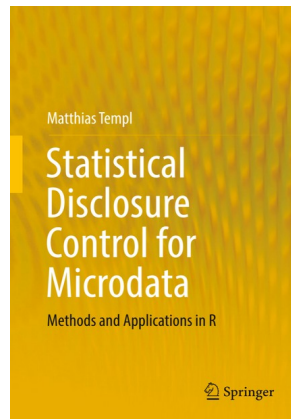
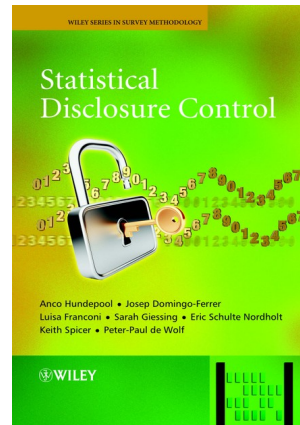
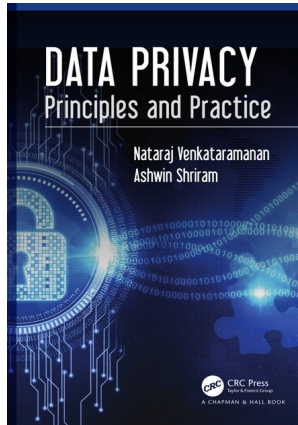
--
-- DP-query with epsilon = 1 and sensitivity = M

select * from (
  select anon_count(henkilotunnus, 1.0/5) as N, tutkimus_nimike, tulos_teksti
  from p_sample
)
where N <= M
```

The Fundamentals of Data Protection

Four levels of data protection:

- 1) Laws and contracts (legal uses, licenses, terms and conditions)
- 2) Organizatorial data management rules (data access control and logging, standard operational procedures)
- 3) Cybersecurity (data encryption, user identification and authentication, digital key and certificate management)
- 4) Statistical Disclosure Control** (different data anonymization methods)



There are many books written about Statistical Disclosure Control, but I personally suggest reading the introductory section of Damien Desfontaines PhD Thesis “Lowering the cost of anonymization” (2020) available at: <https://desfontain.es/thesis/>



Protecting Sensitive Unit-Level Data

Personal
information

1. **Original data:** Individual, up-to-date personal information directly from the source systems
2. **De-identified data:** Data protected with pseudonym. Direct identifiers, such as personal identification numbers are changed to codes (e.g. P001, P002, P003...)
3. **Non-anonymous synthetic data:** Bootstrap re-sampling, non-anonymous GAN models,...



Not personal information
Residual risk
No residual risk

4. **Anonymous synthetic data:** Data which is (i) derived from real data, is (ii) anonymized and (iii) has similar statistical properties with the original real data.
5. **Anonymous data:** De-identified (and possibly otherwise processed data) from which it is practically impossible to re-identify the person behind a given record. The identification key has also been disposed of. If the id-mapping table is stored, we call the data - according to GDPR - **pseudonymous**.
6. **Simulated data:** Data generated from a mathematical model (such as a mechanistic differential equation) with statistical properties similar to the real one. (e.g. Covid 19 pandemic forecasts)
7. **Fabricated data:** E.g. correctly formatted email addresses and personal IDs for software testing. No statistical predictive power.

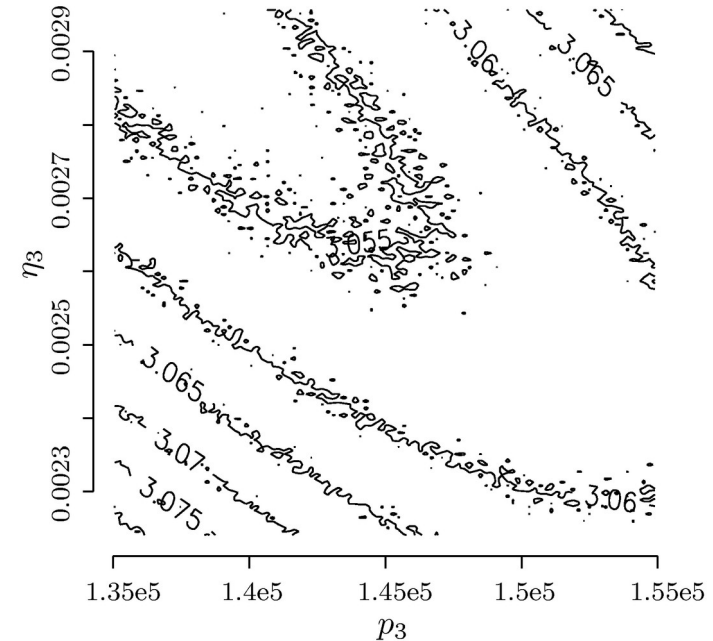


Contemporary Methods for Data Protection

1. **Employ meta-analyses** (i.e. combine data on aggregated, statistical level). Not always an option – especially if we need to match individual row data to compute the statistics.
2. **Copy data to one trusted register keeper** or public authority, such as [Findata](#), and apply strict security policy on data access.
3. **Anonymize data before release.** Unfortunately, this usually implies the removal and heavy aggregation of attributes or adding lots of noise.

Future Trends?

- **Federated Computing.** Distributed data, unified computing.
- **Synthetic Data.** Finds ways to generate realistic, but completely artificial data through non-parametric statistical modelling



Simulated data with artificial noise for algorithm testing.



Motivation for Anonymous Synthetic Data

Real data might be unavailable, because

- the use case is incompatible with the security requirements, or
- the data does not exist yet.

In the absence of real data, one can *fabricate data*, for example by running a dynamical simulation model.

Fabricated data makes sense in ICT development, where e.g. user interfaces need to be tested with realistic-looking prototype data.

Nonetheless, e.g. for hackathons, we need more realistic data.



Data Anonymization Methods

According to Damien Defontaines [1], data protection methods can be divided into *syntactic* and *semantic* methods.

- 1) Syntactic methods:** Use minimum frequencies (k-anonymity) and other metrics like l-diversity and masking to protect the individual records
- 2) Semantic methods:** Use Differential Privacy or other methods that concentrate on the *anonymity of the data generation process* (e.g. database query), not on a particular output.

Syntactic methods are rather easy to understand and apply (similar to Bohr's atomic model in chemistry), but they suffer from certain paradoxes if scrutinized further. For example, we might get records that match real living (or yet unborn) people just by running a simulator or using a simple random number generator. This dilemma is solved by requiring that instead of concentrating the output, the *process of data generation* should guarantee anonymity.

Unfortunately, Differential Privacy (the quantum mechanics of anonymization?) is not an explicit algorithm, but a criterion. One still needs to invent algorithm that satisfy DP, such as frequency counts.

[1] See Damien Desfontaines, Lowering the cost of anonymization, 2020, ETH Zurich. PhD Thesis (<https://desfontain.es/thesis/>)



Examples of Synthetic Data



	A	B	C	D	E	F	G	H	
1	henkilötunnus	resepti_numero	atc_koodi	atc_selite	laakeaine	laakeainetyyppi	kauppanimi	aika_pvm	annostelu
2	191245-WOFM	94846433232635478	B01AB05	Enoksapariini	enoksapariininatrium	vaikuttava_aine	KLEXANE	25.11.2014	1 tabletti 3 kertaa vuorokaudessa 7 vrk ajan. Antibiootti.
3	211071-ER4K	016062498022959001	L02BA01	Tamoksifeeni	tamoksifeenisitraatti	vaikuttava_aine	TAMOFEN	14.3.2017	1 tabletti päivässä
4	221147-1NWW	516297952406825298	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	24.11.2013	1 tabletti x1-3.
5	220246-XJAW	390581673860859896	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	12.5.2015	1 - 2 annosta aamuin illoin, suu huuhdotaan lääkkeen oton jälkeen
6	160536-V98T	468936691601333898	S01XA20	Keinokyyneleet ja muut luokittelemattomat valmisteet	hypromelloori	vaikuttava_aine	HYPROSAN	5.4.2016	1 tabletti tarvittaessa 1-3 kertaa vuorokaudessa.
7	260378-ATDY	934646845134571086	D10BA01		isotretinoini	vaikuttava_aine	ISOTRETINOIN ACTAVIS	23.1.2014	2 tabl iltaisin
8	221254-GIQ6	703892520511949149	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	15.7.2015	1 tabletti aamuisin.
9	040579-8CZN	25222573573643988	J07AL02	Pneumokokkikonjugaattirokote	pneumokokkipolysakkaridi, konjugoitu	vaikuttava_aine	PREVENAR 13	10.6.2013	1*3 tarv. särkyyn.
10	060960-9XP1	376039765490494225	J01CE02	Fenoksimetyylipenisilliini	fenoksimetyylipenisilliinikalium	vaikuttava_aine	V-PEN 1500	20.4.2016	1 tabl. 2:sti päivässä tulehduksen hoitoon.
11	190646-WOV3	055494077593879689	A04AA02	Granisetroni	granisetronihydrokloridi	vaikuttava_aine	GRANISETRON STADA	14.9.2014	1annos aamuin illoin astmaan. Pahanemisvaiheissa 2+2 1-2 viikkoa.
12	241228-TV8R	590469795430918820	H03AA01		levotyrokisiinatrium vastaten levotyrokisina	vaikuttava_aine	THYROXIN	27.5.2017	1 tabl tarv. 1-3 kertaa päivässä. Kipulääke.
13	180690-9TSM	292077063455582840			MELATONIN, melatoniini, 3 mg, -	muu_laake_tiedot		13.2.2013	1 tabl iltaisin. Mielialalääke.
14	300953-SU8H	152904720270561320	H02AB06	Prednisoloni	prednisoloni	vaikuttava_aine	PREDNISOLON	7.6.2013	1 tabletti kerran päivässä.
15	250296-Y52F	769604541680079562	C07AB07	Bisoprololi	bisoprololifumaraatti	vaikuttava_aine	BISOPROLOL RATIOPH	29.8.2016	4 ml injektio lihakseen 12 viikon välein.
16	180489-1GIN	049970884043963441	R05FA02	Opiumjohdokset ja ekspektorantit, lukuun ottamatta m	kodeiini-fosfaatti-hemihydraatti,efedriinihydrokloridi,dit	vaikuttava_aine	CODESAN COMP	10.7.2015	1 tabl tarvittaessa kerran päivässä
17	270462-8FHK	969806661300809769	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	25.11.2014	D.S. 1 annos mol. sieraimen 1-2:sti päivässä. Tukkoisuuteen.
18	220334-A56K	284347271920969936	G03GA05	Folliotropiinalfa	folliotropiini alfa	vaikuttava_aine	GONAL-F	17.9.2012	1 tbl 1-3 kertaa vrk:ssa kipuun.
19	240928-F2MA	615408308811748728	R06AE09		levosetiritsiini-dihydrokloridi	vaikuttava_aine	XYZAL	10.12.2012	1 tabletti 2 kertaa päivässä.
20	270597-DYZA	996839916211359869	S01BA04		prednisoloni-asetaatti	vaikuttava_aine	PRED FORTE	29.10.2015	1 tabletti kerran päivässä
21	031228-P947	470962128725546216	M01AB05		diklofenaakkinaatrium	vaikuttava_aine	MOTIFENE DUAL	31.1.2013	1 tabl 1-2 kertaa päivässä, pahoinvoinnin hoitoon. Ja ohjeen mukaan sy
22	211173-ZC9S	511785582179482718	N06BA04	Metyylifenidaatti	metyylifenidaatti-hydrokloridi	vaikuttava_aine	CONCERTA	10.6.2017	Erillinen ohje, keltarauhashormoni
23	180673-CVBU	070559495848135649	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	2.6.2017	1 tabl 1-3sti päivässä särkylääke
24	101064-JTAR	101304519953659541	C10AA01		simvastatiini	vaikuttava_aine	SIMVASTATIN ACTAVIS	3.6.2011	1 tabl tarv 3x/pv kipuun
25	080336-UEF8	085731549209808924	L02AE02	Leuporeliini	leuporeliini-asetaatti	vaikuttava_aine	ENANTON DEPOT DUA	9.1.2016	1 pussi/vrk
26	040152-MR3H	790250060740349210	C07AB02		metoprololi	vaikuttava_aine	SPECICOR DOS	11.2.2017	Yksi tabletti kerran päivässä
27	150693-OBTY	345439526087897692	A02BC02		pantopratsolnatrium-seskvihydraatti	vaikuttava_aine	SOMAC	23.8.2012	1 tabl päivässä diabeteslääke
28	171077-TMX8	802194387744663500	N05AN01	Litium	litiumkarbonaatti	vaikuttava_aine	LITO	30.10.2016	1 tabletti x1. kahdesti vuodessa kuukauden tauko.
29	290359-1EAK	960209612536429827	N06AB10	Essitalopraami	essitalopraamioksalaaatti	vaikuttava_aine	ESCITALOPRAM ORION	1.12.2014	D.S. Kipulääke1 tabletti kaksi kertaa päivässä 5 päivääTämän jälkeen 1 t
30	260735-2NBH	209913678314706540	A02BC02		pantopratsoli	vaikuttava_aine	PANTOPRAZOL ACTAV	7.11.2014	1 tabletti 1-3 kertaa päivässä tarvittaessa kipuun.
31	110988-FZ7X	197178901574453114	N02AA05	Oksikodoni	oksikodonihydrokloridi	vaikuttava_aine	OXYNORM	9.5.2011	1 tabletti 3 kertaa päivässä 5 päivän ajan.
32	080553-XAJI	074706311575004863	H02AB06	Prednisoloni	prednisoloni	vaikuttava_aine	PREDNISOLON	21.8.2015	1 tabl 3 kertaa päivässä divertikuliittiin. Kuuri loppuun
33	180178-4GBX	843818990382363714			AQUALAN	kauppanimi	AQUALAN	12.6.2014	D.s. 1 tabletti 2 kertaa viikossa emättimeen, paikallinen estrogeenihoito
34	180259-UW69	241336540479470602			Tabl Acid Folic 5 mg	muu_laake_tiedot		13.8.2014	1x2.
35	160706AXKBE	037628154449967810	M01AE01	Ibuprofeeni	ibuprofeeni	vaikuttava_aine	BURANA	14.3.2016	1 tabl 1-3 kertaa päivässä tarvittaessa kipuun
36	020435-V1KC	284817460573826452	N02BE01	Parasetamoli	parasetamoli	vaikuttava_aine	PANADOL FORTE	13.7.2013	1 tabletti tarvittaessa 3 kertaa päivässä
37	251255-DC5I	989378565818591659	N07XX09	Dimetyylifumaraatti	dimetyylifumaraatti	vaikuttava_aine	TECFIDERA	3.4.2017	D5. 1 tabl tarvittaessa 4x/vrk särkyyn
38	120949-MACI	117072610705513568	S01CA01		deksametasoni, kloramfenikoli	vaikuttava_aine	OFTAN DEXA-CHLORA	13.3.2015	1 tabletti 1-3 kertaa päivässä tarvittaessa kipuun.
	resepti	toimenpide	laakemaarays	opera_leikkaus_toimenpide	radu_vsshp	palvelu_diagnoosi	diagnoosi	teksti	asiakas
	labra_vs ...								

Enter a part of object name here

- stage_ipana_ensemble
- stage_labdw
- stage_medikro
- stage_medikro_2
- stage_musti
- stage_mynla
- stage_potilaspaalaut
- stage_qpata
- stage_resmed
- stage_sai
- stage_sasdata
- stage_sentrysuite
- stage_silmaklinikka
- stage_uraods
- synthetic
 - Tables
 - asiakas 1.6M
 - diagnoosi 2.5M
 - laakemaarays 2.5M
 - labra_vsshp 28M
 - opera_leikkaus_toimenpide
 - palvelu_diagnoosi 3.8M
 - radu_vsshp 2.4M
 - resepti 2.4M
 - teksti 2.3M
 - toimenpide 20M
 - upload 72K
 - Views
 - Materialized Views
 - Indexes
 - Functions
 - Sequences
 - Data types
 - Aggregate functions
 - syoparekisteri
 - tekstitykki

Project - General x

Name DataSource

- Bookmarks
- Diagrams
- Scripts

ER Diagram

upload	
144	upload_id
☒	is_uploaded
abc	source_schema
abc	source_table
abc	target_schema
abc	target_table
abc	column_list
abc	criteria_list
abc	table_creation_script
abc	last_load
abc	current_load

toimenpide	
abc	henkilotunnus
abc	tyyppi
abc	palvelu_numero
123	rivi_numero
abc	palvelu_tunnus
abc	pot_eala_koodi
abc	pot_eala_selite
abc	kayntityyppi_koodi
abc	kayntityyppi_selite
abc	toimenpide_koodi
abc	toimenpide_selite

asiakas	
abc	henkilotunnus
abc	potilasnumero
abc	yhd_viite
abc	sukupuoli_selite
abc	syntymaika_pvm
abc	kuolin aika_pvm
abc	sukunimi
abc	etunimi
abc	jakeluosoite
abc	column_list
abc	ammatti
abc	aidinkieli_selite
abc	pot_kotikunta_selite
abc	pot_kotikunta_koodi
abc	vrk_kotikunta_selite
abc	vrk_kotikunta_koodi
abc	ominaisuus1
abc	tapahtuma_aikaleima

labra_vsshp	
abc	henkilotunnus
123	labra_id
abc	tutkimusrivivain
abc	naytetunnus
abc	putkien_lkm
abc	nayteenotto
abc	paatutkimuspaketti
abc	tutkimus
abc	hoitava_laitos_koodi
abc	tutkimuksen_tyyppi
abc	tulos_teksti
abc	tuloksen_mittayksikko
abc	tarkein_mikrobi
123	viitearvo_min
123	viitearvo_max

resepti	
abc	henkilotunnus
abc	resepti_numero
abc	atc_koodi
abc	atc_selite
abc	laakeaine
abc	laakeainetyyppi
abc	kauppanimi
abc	aika_pvm
abc	annostelu
abc	annos
123	pakkausko_kokonaismaara
123	pakkaus_lkm
123	kesto aika
abc	kesto aika_yksikko_selite
abc	vahvuus
abc	muoto
abc	paivittaja_yksikko_nimi
abc	tapahtuma_aikaleima

opera_leikkaus_toimenpide	
abc	henkilotunnus
123	pk_dmf_leikkaus
123	leikkaus_tmp_act_id
abc	hoitopaatospvm
abc	varaus_pvm
abc	suunniteltu_pvm
abc	toteutunut_pvm
abc	pituus
abc	paino
abc	asa_luokka
abc	potilaan_erikoisala_koodi
abc	potilaan_erikoisala_nimi
abc	hoitava_osasto_koodi
abc	hoitava_osasto_nimi
abc	hoitava_laitos_koodi
abc	hoitava_laitos_nimi
abc	toimenpideosasto_koodi
abc	toimenpideosasto_nimi
abc	lahettava_osasto_koodi
abc	lahettava_osasto_nimi
abc	jatkohoitosasto_koodi

diagnoosi	
abc	henkilotunnus
abc	dgn_pvm
123	ika
abc	palvelu_numero
abc	dgn_numero
abc	diagnoosi
☒	on_syy
☒	on_tapahtuma_dgn
abc	selite
abc	paadgn
abc	tarkenne1
abc	tarkenne2
abc	vo_toimipiste_koodi
abc	yksikko_nimi
abc	tulosyksikko_selite
abc	vastuualue
abc	pot_eala_koodi
abc	pot_eala_selite
abc	tapahtuma_aikaleima

palvelu_diagnoosi	
abc	henkilotunnus
abc	potilasnumero
abc	palvelu_numero
abc	varaus_numero
abc	varaushetki_pvm
abc	kayntityyppi_selite
abc	kayntityyppi_tarkenne_selite
abc	alku hetki_pvm
abc	alku hetki_min
abc	loppu hetki_pvm
abc	loppu hetki_min
abc	vo_toimipiste_koodi
abc	yksikko_nimi
abc	kustannuspaikka
abc	hoi_laakari_koodi
abc	hoi_laakari_nimi
abc	hoi_laakari_yksiloitintunnus
abc	pot_eala_koodi
abc	pot_eala_selite
abc	res_koodi
abc	res_selite

teksti	
abc	henkilotunnus
abc	potilasnumero
abc	teksti_muoto
abc	hoitotapahtuma_numero
abc	hoitotapahtuma_tunnus
abc	hoitotapahtuma_alku hetki
abc	hoitotapahtuma_alku hetki_arvio
abc	hoitotapahtuma_loppu hetki
abc	potilaskertomus_numero
abc	toimipiste_koodi
abc	toimipiste_nimi
abc	yksikko_nimi
abc	nakyma_selite
abc	tila
abc	teksti_numero
abc	teksti
abc	kertomus_paivityshetki_s
abc	kertomus_paivityshetki_ods
abc	teksti_paivityshetki_ods
abc	teksti_aika
abc	paivityshetki_mine

laakemaarays	
abc	henkilotunnus
abc	laakemaarays_numero
abc	kauppanimi
abc	vaikuttava_aine
abc	atc_koodi
abc	vahvuus
abc	muoto
abc	antotapa_koodi
abc	antotapa_selite
123	kotilaake
123	n_laake
123	antibiootti
123	ex_tempore
abc	alkuaika_pvm
abc	alkuaika_min
abc	loppuaika_pvm
abc	loppuaika_min
abc	lopetus_syy_koodi
abc	lopetus_syy_selite
abc	poisto_syy_koodi
abc	poisto_syy_selite

radu_vsshp	
abc	henkilotunnus
123	radulahete_id
123	radututkimus_id
abc	tehty_tutkimus_koodi
abc	tehty_tutkimus_selite
abc	tutkimuksen_tyyppi
abc	tutkimusryhman_selite
abc	tutkimushetki
abc	onko_paivystys
123	tutkimuksen_hinta
abc	lahetehetki
abc	ajanvaraus hetki
abc	tilaava_yksikko
abc	hoitava_yksikko
123	ulkopuolinen_tutkimus
abc	suorittava_toimipiste
abc	potilaan_kotikunta
abc	lausunto_idt
abc	tutkimuksen_ac_numero
123	dw_radututkimus_id
abc	dw_tutkimus_paivityshetki
abc	dw_lahete_paivityshetki
abc	paivityshetki_stage
abc	tapahtuma_aikaleima

11 objects

100%

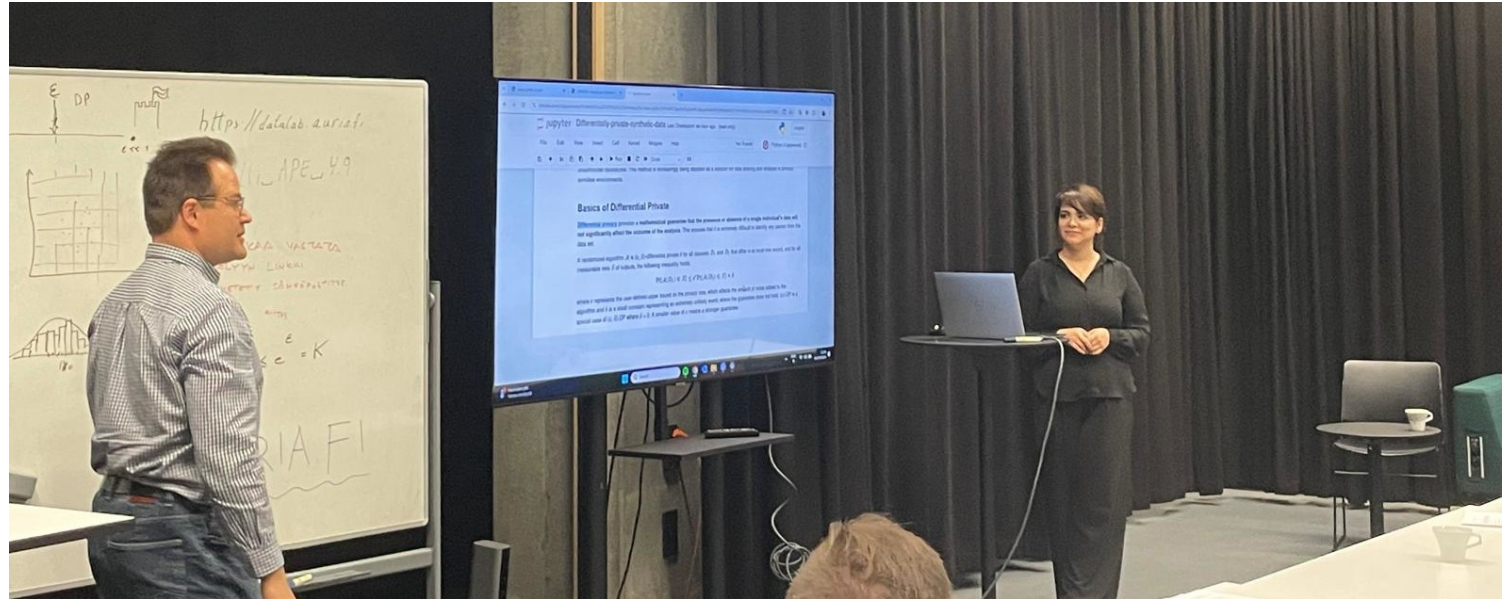
Keep layout Save ... Revert Refresh



How to Produce Anonymous Synthetic Data

Differential Privacy (DP) is the only “bulletproof” anonymization method with a rigorous mathematical proof, but it is hard to apply (beyond frequency counts).

- 1) For tabular data sets (e.g. with fixed measurement time), DP frequency count can be used
- 2) For longitudinal data set (e.g. time series), DP is not readily applicable, but other syntactic methods exist



2) E.g. Brandon Theodorou, Cao Xiao & Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications* volume 14, Article number: 5305 (2023) <https://www.nature.com/articles/s41467-023-41093-0>

1) Image from [Syndate](#) project workshop for companies on Wed 2024-09-04 at Educity Turku. Parisa Movahedi (right) is presenting [AIM](#) (An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data), which is a synthetic data generation method based on DP, marginal distributions and noisy frequency counts.



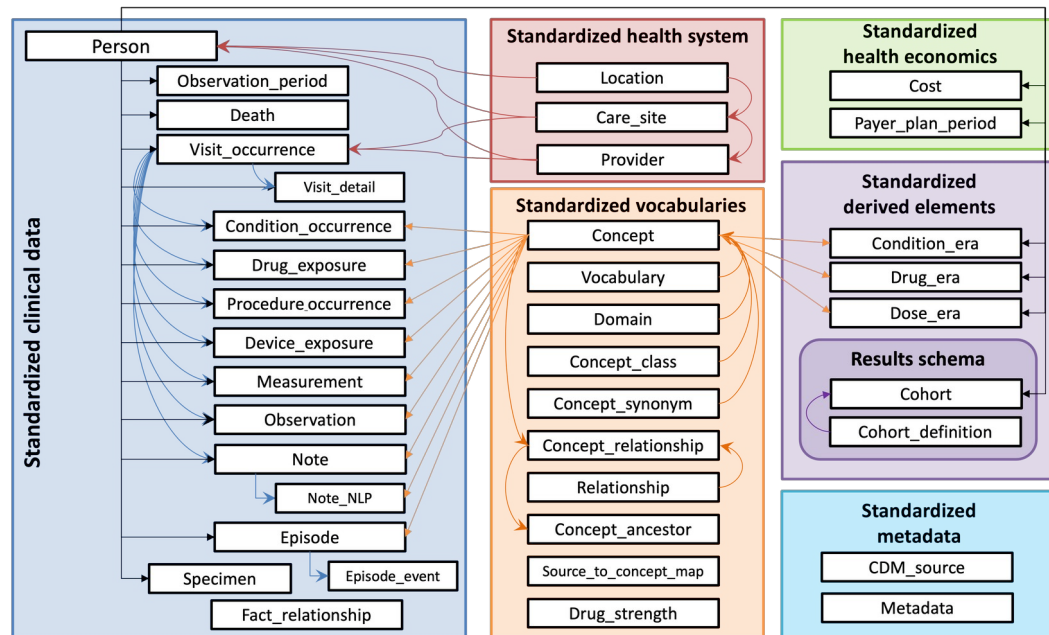
Data Modelling: OMOP Common Data Model

OMOP: Observational Medical Outcomes Partnership.

Was a public-private partnership that was set up to produce information on the use of health care databases to study the effects of medicines.

OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

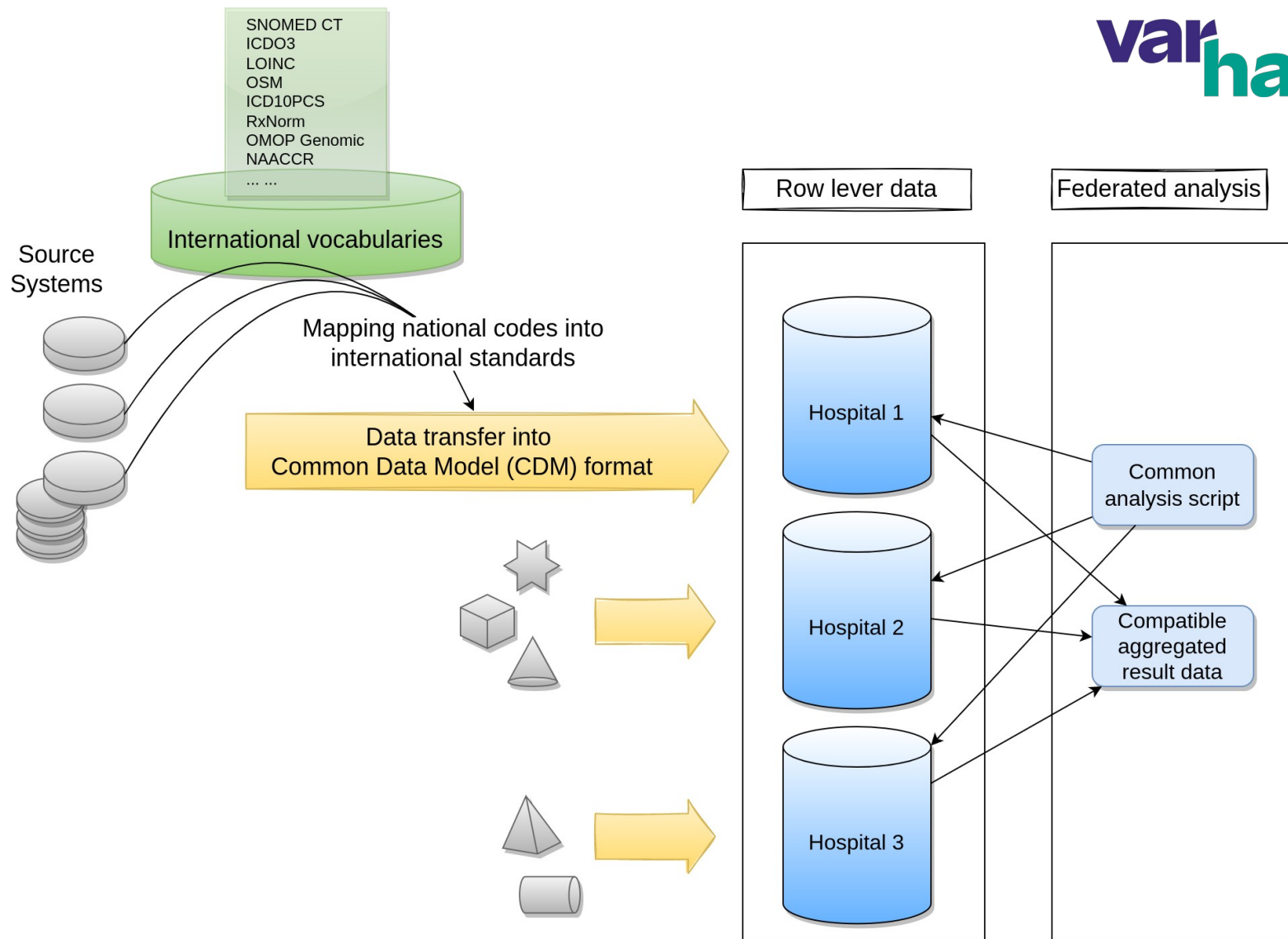
An open community data standard designed to standardize the structure and content of observational data and to enable effective analyses that can produce reliable evidence.



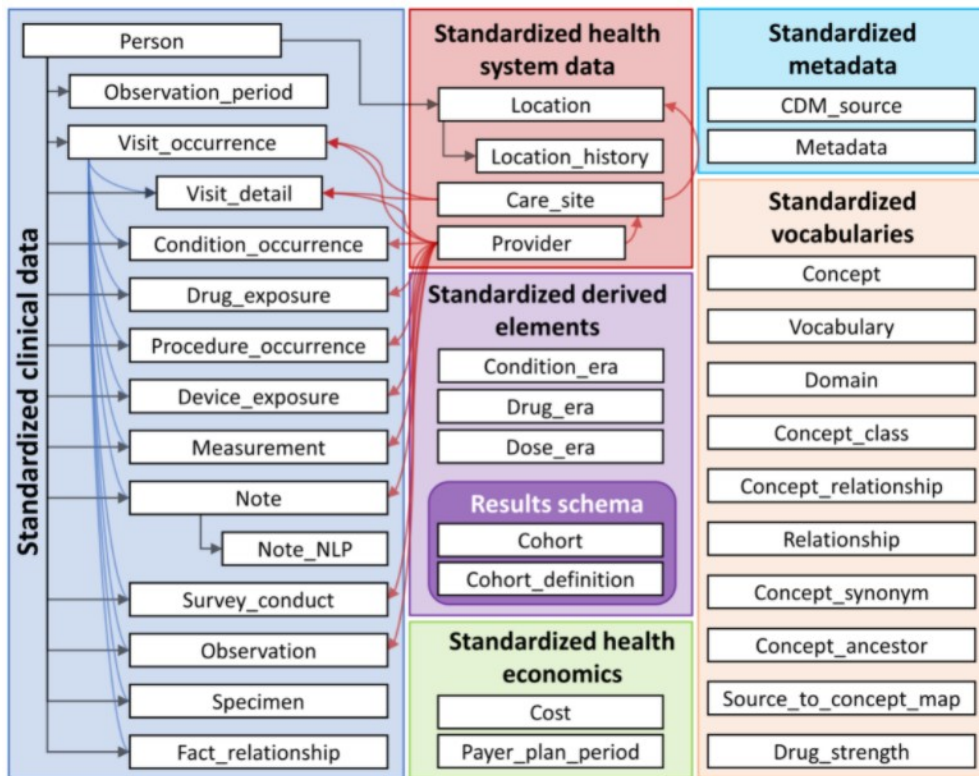
FinOMOP: In 2021, a collaboration of researchers in Finland launched an initiative to create the FinOMOP vocabulary and data model. All university hospitals, the FinnGen study (FIMM, University of Helsinki) and THL participated in it. The work has been funded as EHDEN (European Health Data Evidence Network) projects, and there have also been Finnish small and medium-sized enterprises that have received the EHDEN funding and certificate.



OMOP CDM in Brief



Common Data Model Schema Overview



- The schema is patient-centric (all outcomes are linked to a person)
- Fits well to hospital data, since the model includes all essential concepts

Visit_occurrence	Käynti tai osastohoito
Condition_occurrence	Diagnoosi (kliininen tai patologinen)
Drug_exposure	Lääkitys (määrätty tai annettu)
Procedure_occurrence	Toimenpide
Device_exposure	Laite tai apuväline
Measurement	Laboratoriomittaus
Observation	Muunlainen havainto

Detailed description: <https://ohdsi.github.io/CommonDataModel/cdm53.html>

Data Quality Control

- OHDSI provides several (partially overlapping) tools for quality control, such as
- Data Quality Dashboard, Achilles, CDM Inspection
- These tools can detect
 - Mapping coverage
 - Reference inconsistencies (e.g. diagnoses that do not appear in visit_occurrence table)
 - Infeasible values (e.g. occurrences before birth, male diseases in women, impossible laboratory values)



TYKS JA VARSINAIS-SUOMEN
SAIRAANHOITOPIIRI - TYKS AND
HOSPITAL DISTRICT OF
SOUTHWEST FINLAND

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

TYKS JA VARSINAIS-SUOMEN SAIRAANHOITOPIIRI - TYKS AND HOSPITAL DISTRICT OF SOUTHWEST FINLAND

DataQualityDashboard Version: 1.0.0

Results generated at 2022-01-14 01:49:23 in 7 hours

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1952	27	1979	99%	286	1	287	100%	2238	28	2266	99%
Conformance	565	2	567	100%	78	0	78	100%	643	2	645	100%
Completeness	321	5	326	98%	11	0	11	100%	332	5	337	99%
Total	2838	34	2872	99%	375	1	376	100%	3213	35	3248	99%

- The quality of the transformed data cannot exceed the quality of the original, but the opposite is likely
 - Accuracy is occasionally lost
 - Errors can happen
 - All details cannot be saved in CDM

CDM in Detail

cohort
147 cohort_definition_id
147 subject_id
1 cohort_start_date
1 cohort_end_date

metadata
123 metadata_concept_id
123 metadata_type_concept_id
ABC name
ABC value_as_string
123 value_as_concept_id
1 metadata_date
2 metadata_datetime

cdm_source
ABC cdm_source_name
ABC cdm_source_abbreviation
ABC cdm_holder
ABC source_description
ABC source_documentation_reference
ABC cdm_etl_reference
1 source_release_date
1 cdm_release_date
ABC cdm_version
ABC vocabulary_version

measurement
147 measurement_id
123 person_id
123 measurement_concept_id
1 measurement_date
2 measurement_datetime
ABC measurement_time
123 measurement_type_concept_id
123 operator_concept_id
123 value_as_number
123 value_as_concept_id
123 unit_concept_id
123 range_low
123 range_high
123 provider_id
123 visit_occurrence_id
123 visit_detail_id
ABC measurement_source_value
123 measurement_source_concept_id
ABC unit_source_value
ABC value_source_value

visit_detail
147 visit_detail_id
123 person_id
123 visit_detail_concept_id
1 visit_detail_start_date
2 visit_detail_start_datetime
1 visit_detail_end_date
2 visit_detail_end_datetime
123 visit_detail_type_concept_id

drug_strength
147 drug_concept_id
147 ingredient_concept_id
123 amount_value
123 amount_unit_concept_id
123 numerator_value
123 numerator_unit_concept_id
123 denominator_value
123 denominator_unit_concept_id
123 box_size
1 valid_start_date
1 valid_end_date
ABC invalid_reason

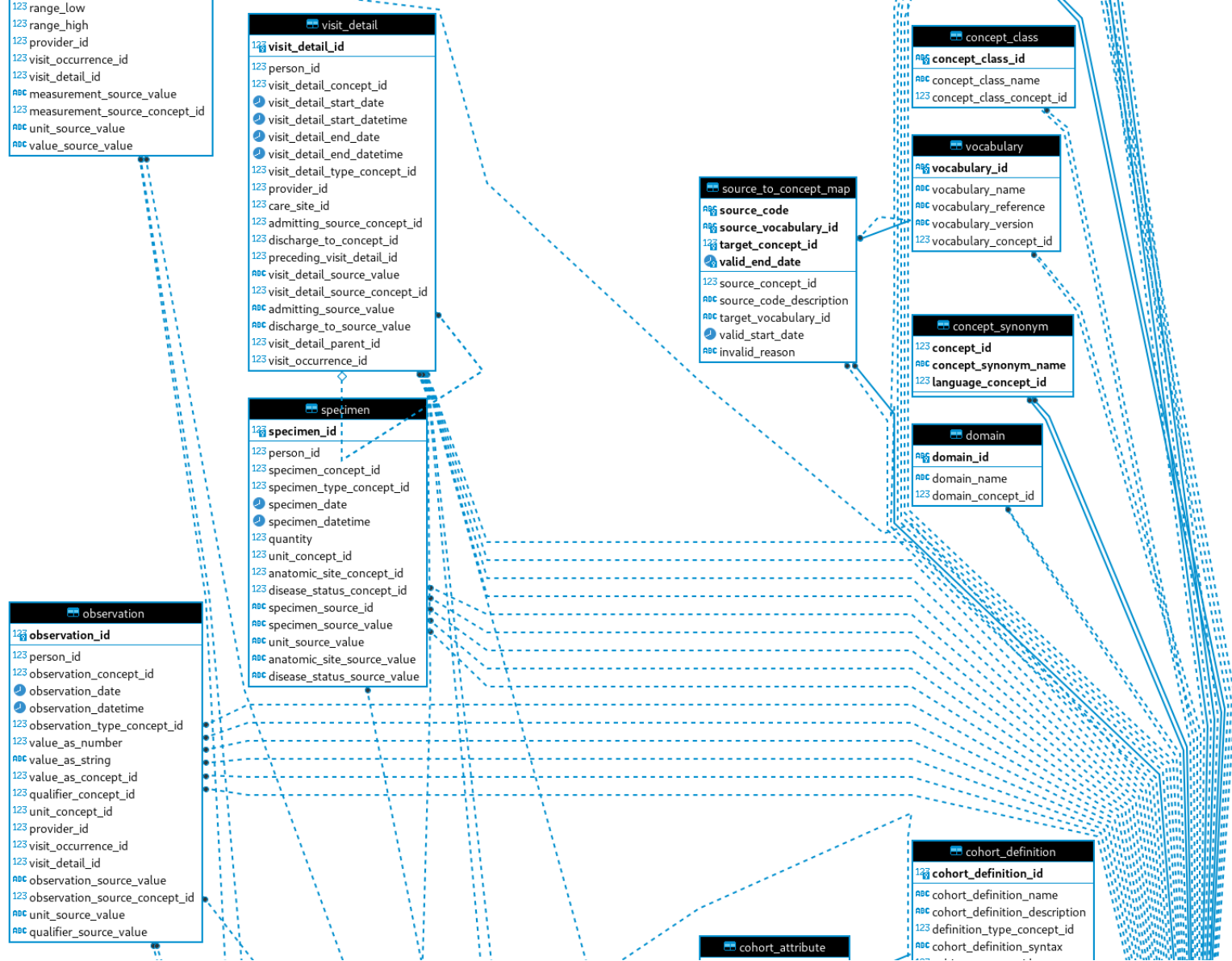
fact_relationship
123 domain_concept_id_1
123 fact_id_1
123 domain_concept_id_2
123 fact_id_2
123 relationship_concept_id

concept_ancestor
147 ancestor_concept_id
147 descendant_concept_id
123 min_levels_of_separation
123 max_levels_of_separation

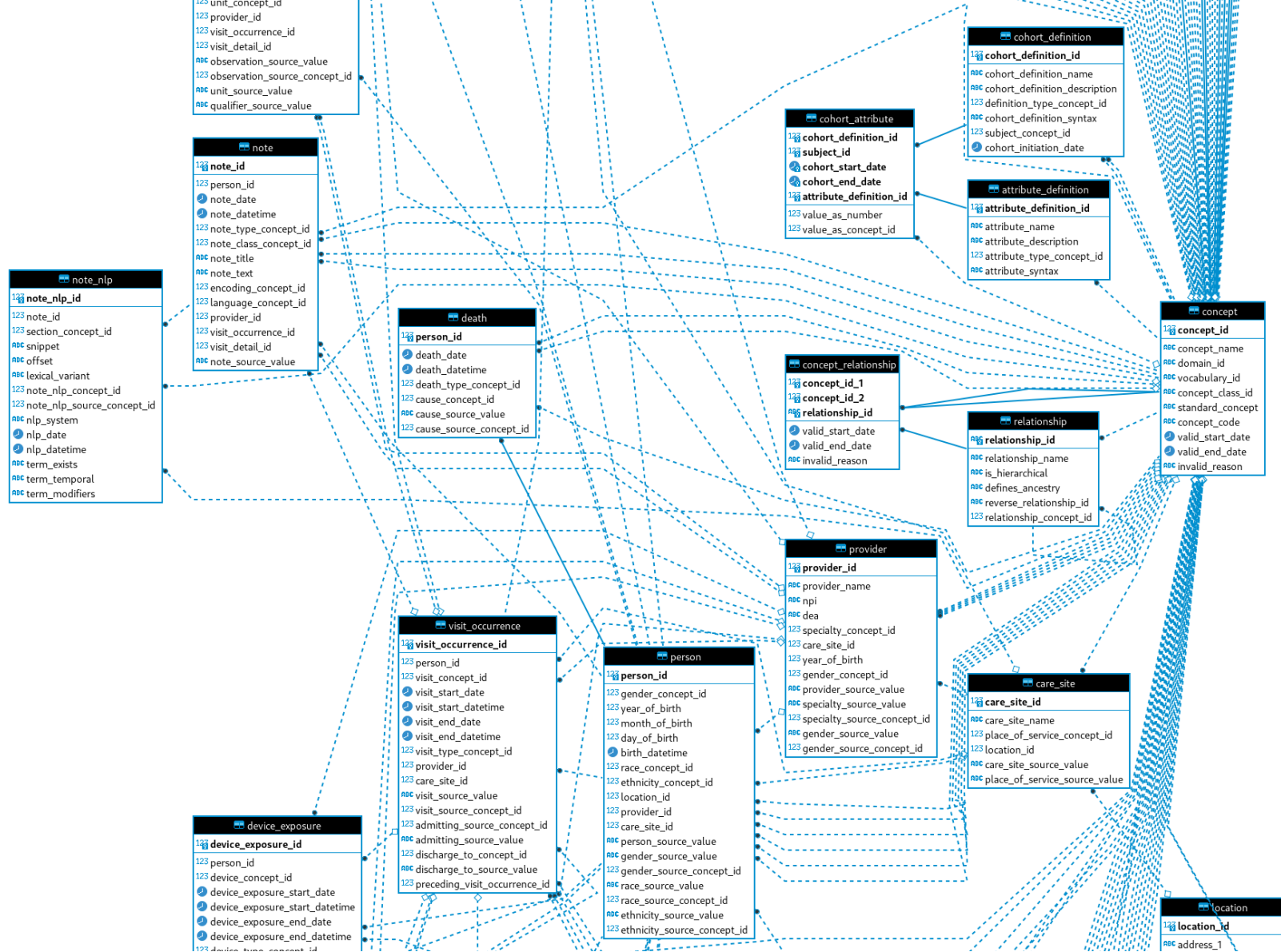
concept_class
ABC concept_class_id
ABC concept_class_name
123 concept_class_concept_id

vocabulary
ABC vocabulary_id

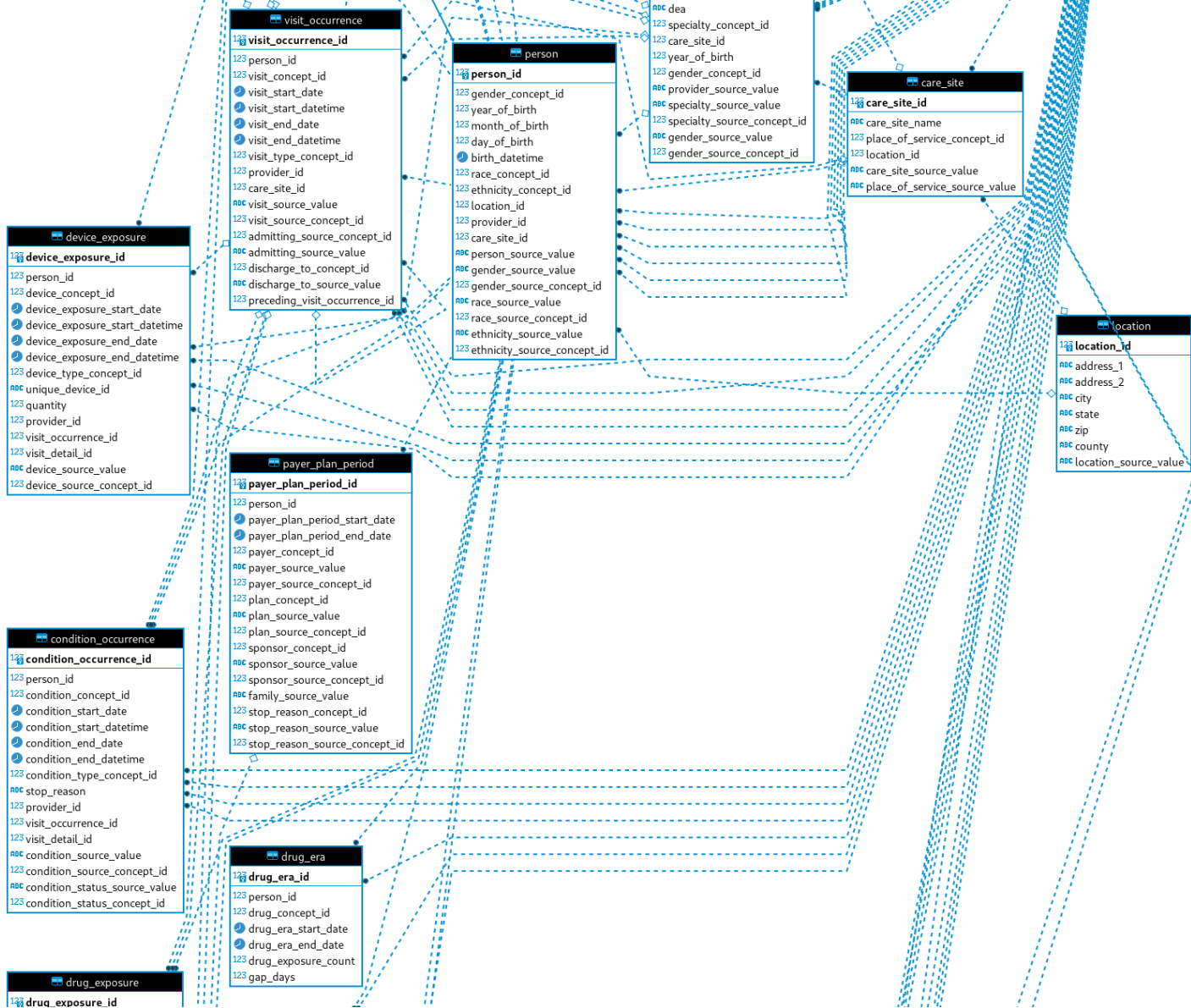
CDM in Detail



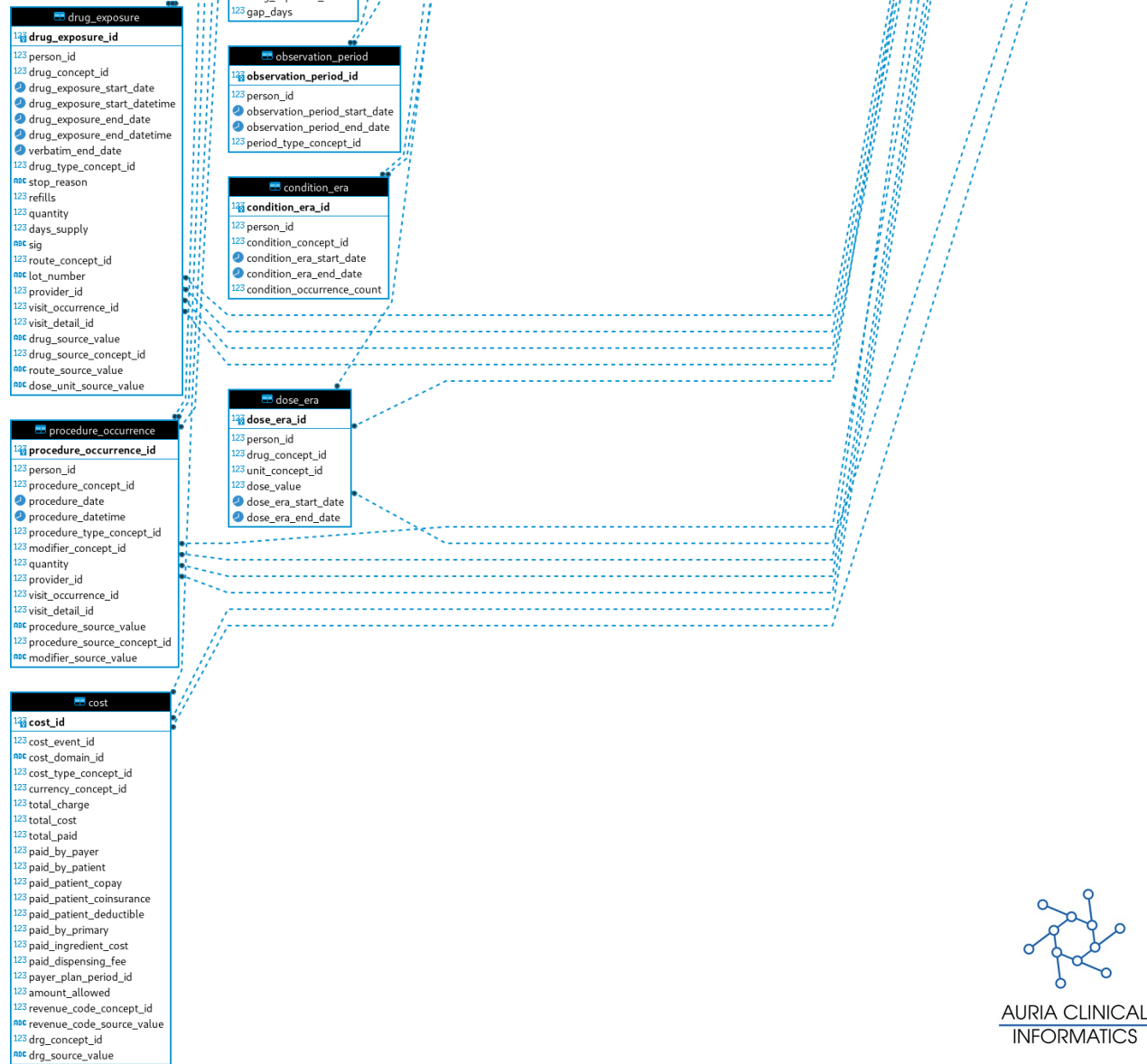
CDM in Detail



CDM in Detail



CDM in Detail



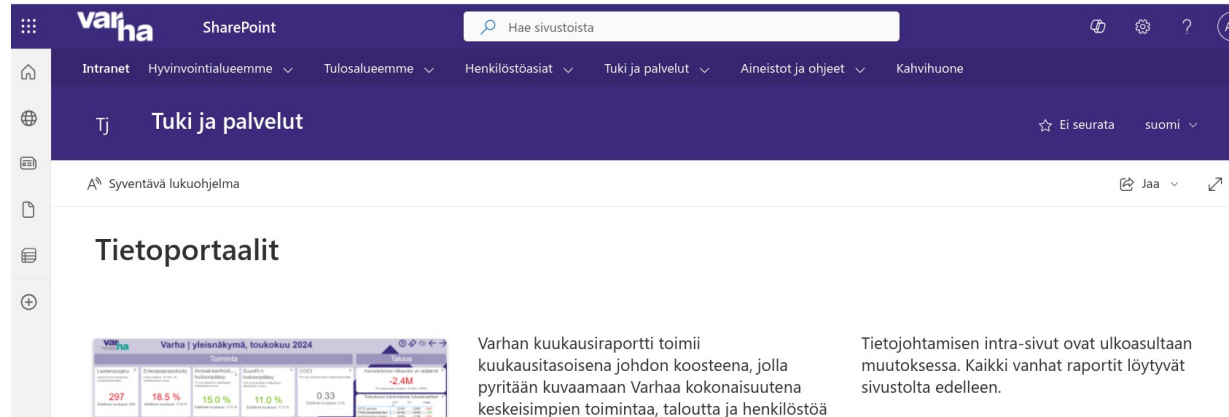
Examples of Data Use

First, some official reports from Varha Intranet

There are two languages to choose, Finnish and Swedish.

There are two separate data teams:

- Management's Data Services (the official figures)
- Experimental & Scientific Data Services (Auria Clinical Informatics)



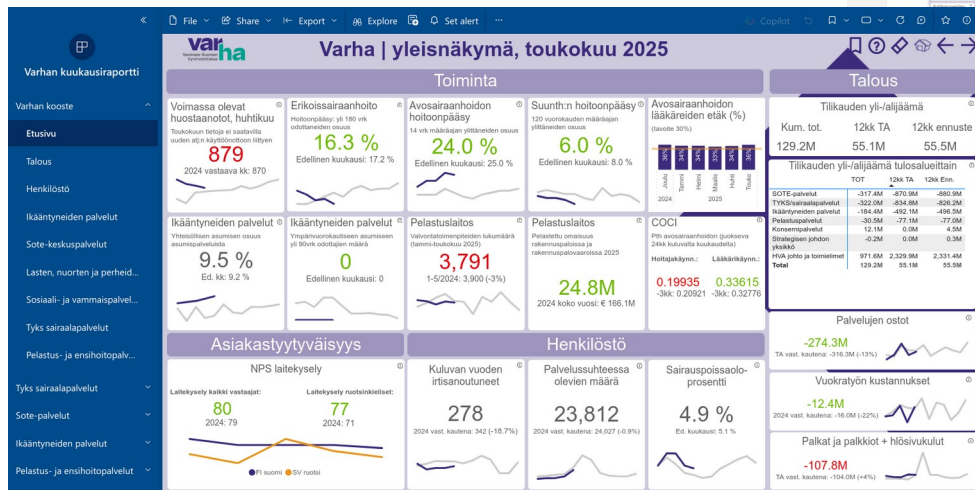
Varhan kuukausiraportti toimii kuukausitasoisena johdon koosteena, jolla pyritään kuvaamaan Varhaa kokonaisuutena keskeisimpien toimintaa, taloutta ja henkilöstöä kuvaavien mittarien avulla.

Tietojohtamisen intra-sivut ovat ulkoiseltaan muutoksessa. Kaikki vanhat raportit löytyvät sivustolta edelleen.

Tietotuote päivittyy edellisen kuun luvuilla aina kuluvan kuun 22. päivä (sen osuessa viikonloppuun tai arkipyhään tehdään päivitys edellisenä arkipäivänä).

Tee tietopalvelupyyntö

Ohjeet



Tietotuotteet



Yhteiset tietotuotteet



Talouden tietotuotteet

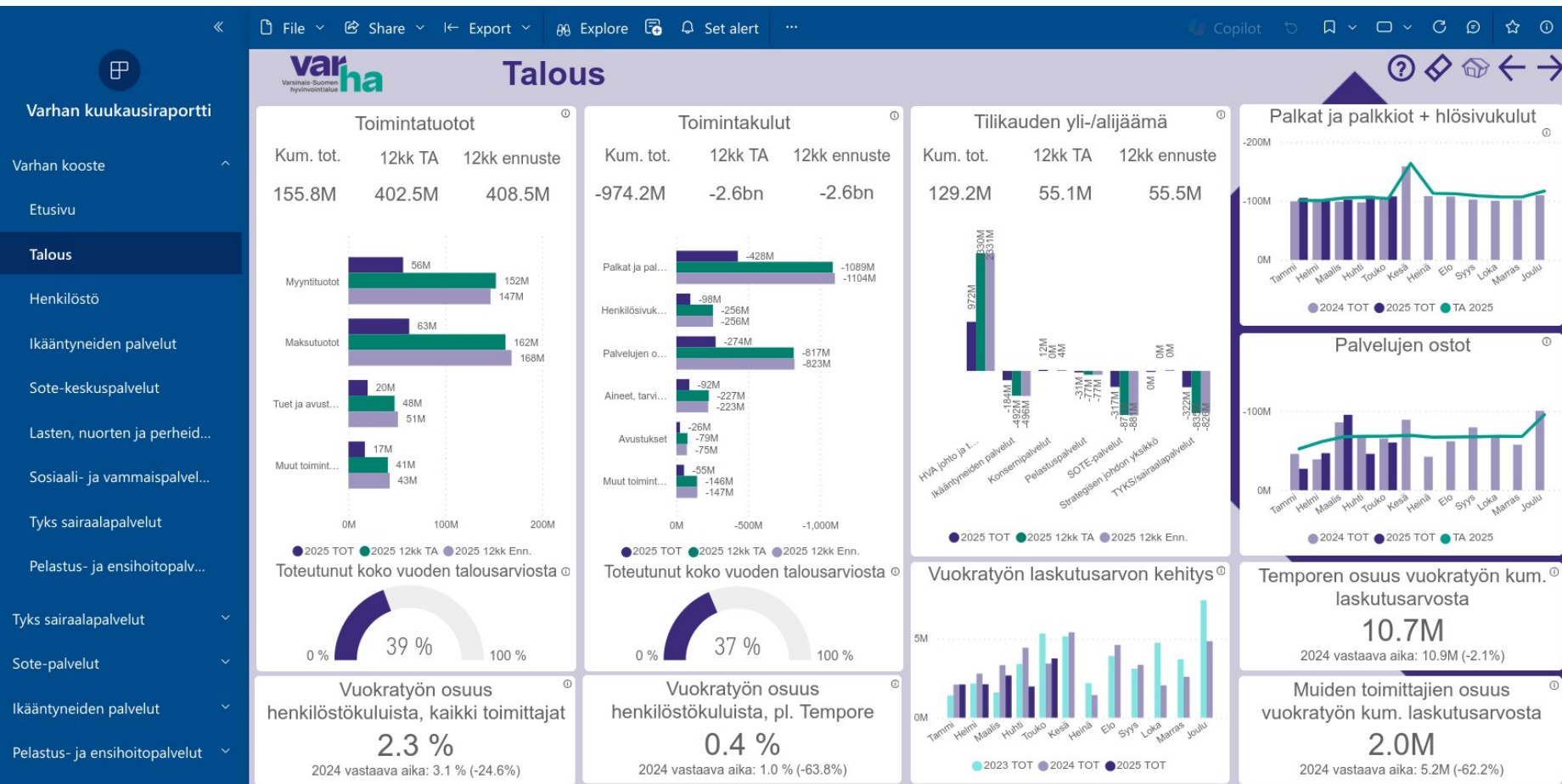


Henkilöstöä koskevat tietotuotteet



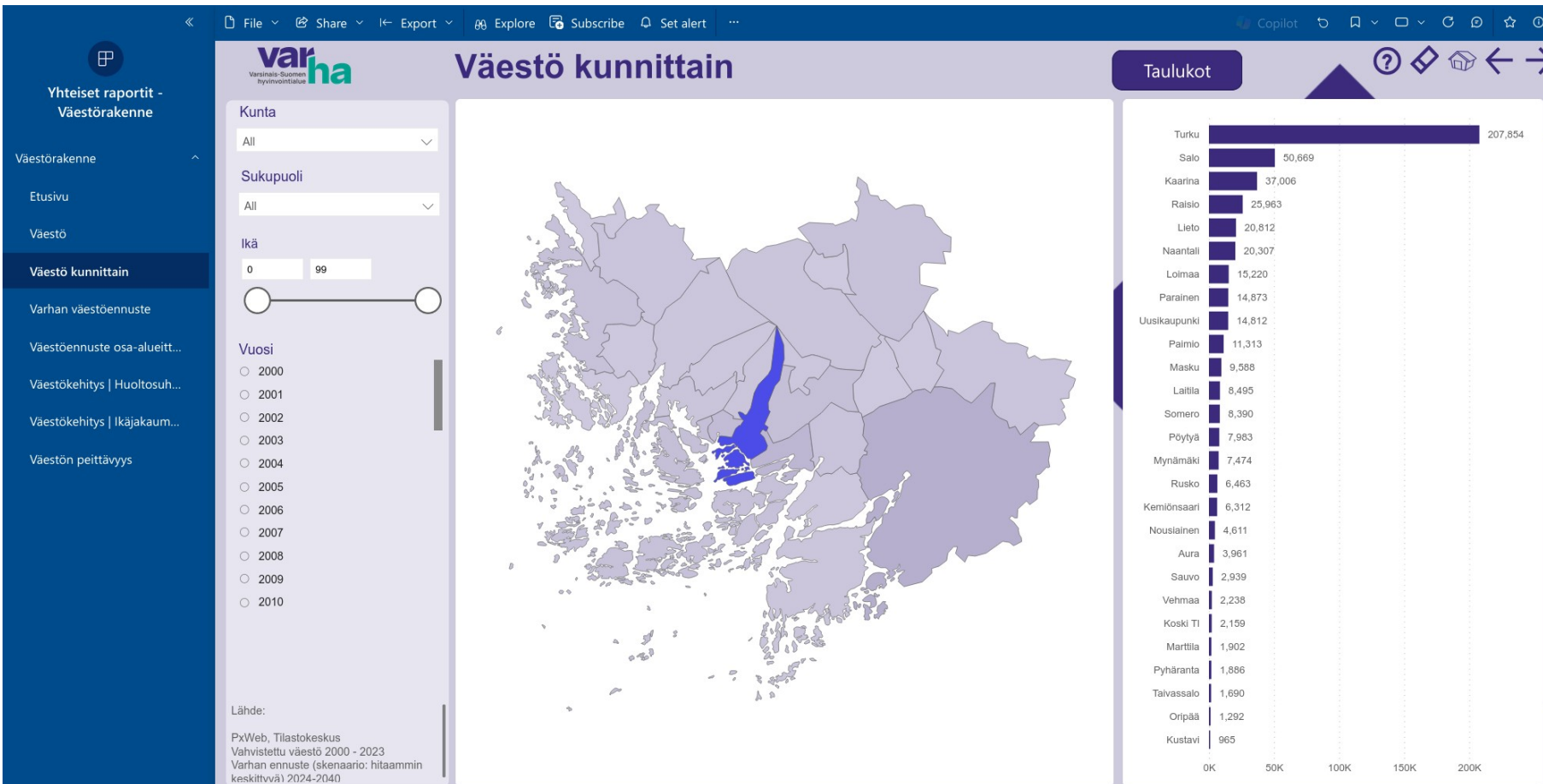
AURIA CLINICAL
INFORMATICS

Examples of Data Use



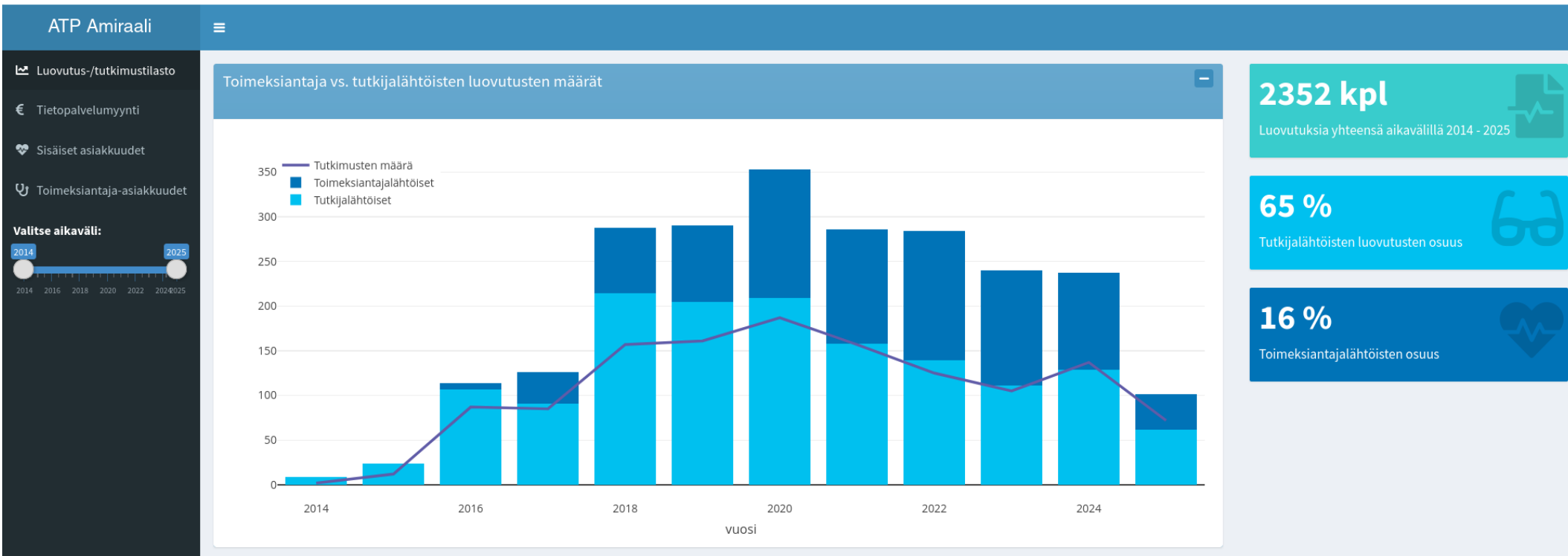


Examples of Data Use



Examples of Data Use

Then, some experimental reports from Auria Clinical Informatics



Unit's Data Export report with R/Shiny + ShinyProxy



Examples of Data Use

Superset - Mozilla Firefox

Superset

https://kide.vsshp.net/login/

AURIA
TETOPALVELU

Sign In

Enter your login and password below:

Username:

arho

Password:

Sign In

Register



Amyotrophic lateral sclerosis (ALS) dashboard with Apache Superset
<https://superset.apache.org/>

Examples of Data Use

Urology Quality Register

with R + Shiny (GPL) +
ShinyProxy (Apache 2) +
AD-authentication (Microsoft)

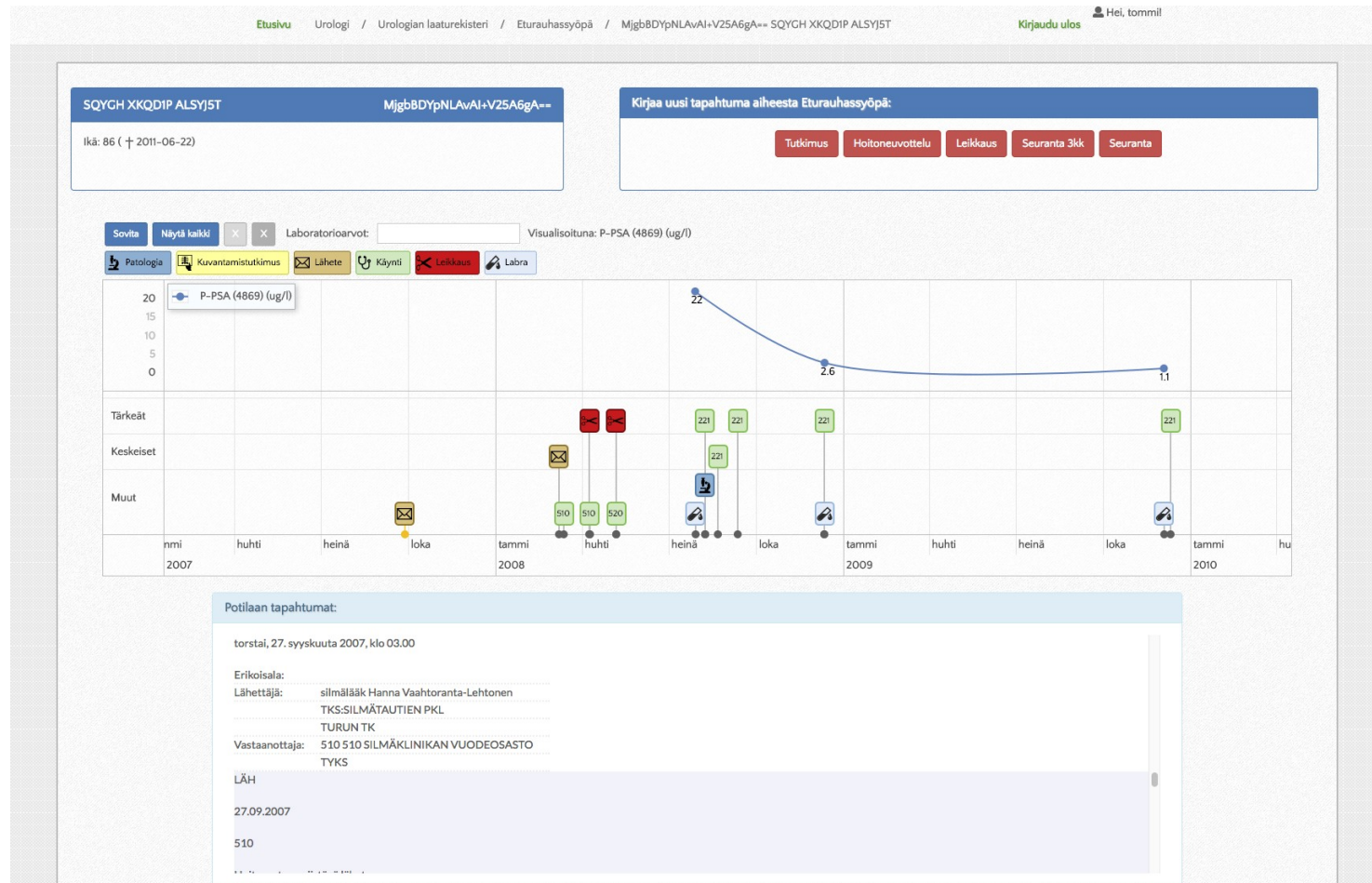


Examples of Data Use

Patient timeline (in use since 2019) for quality control in urology

Implementation: "Full Stack" (PostgreSQL, Flask/Python, D3.js, Bootstrap, ...)

Lessons learned: For better maintainability, use reporting tools, do not program from scratch.



Examples of Data Use

One of the earliest visualizations (2014!)

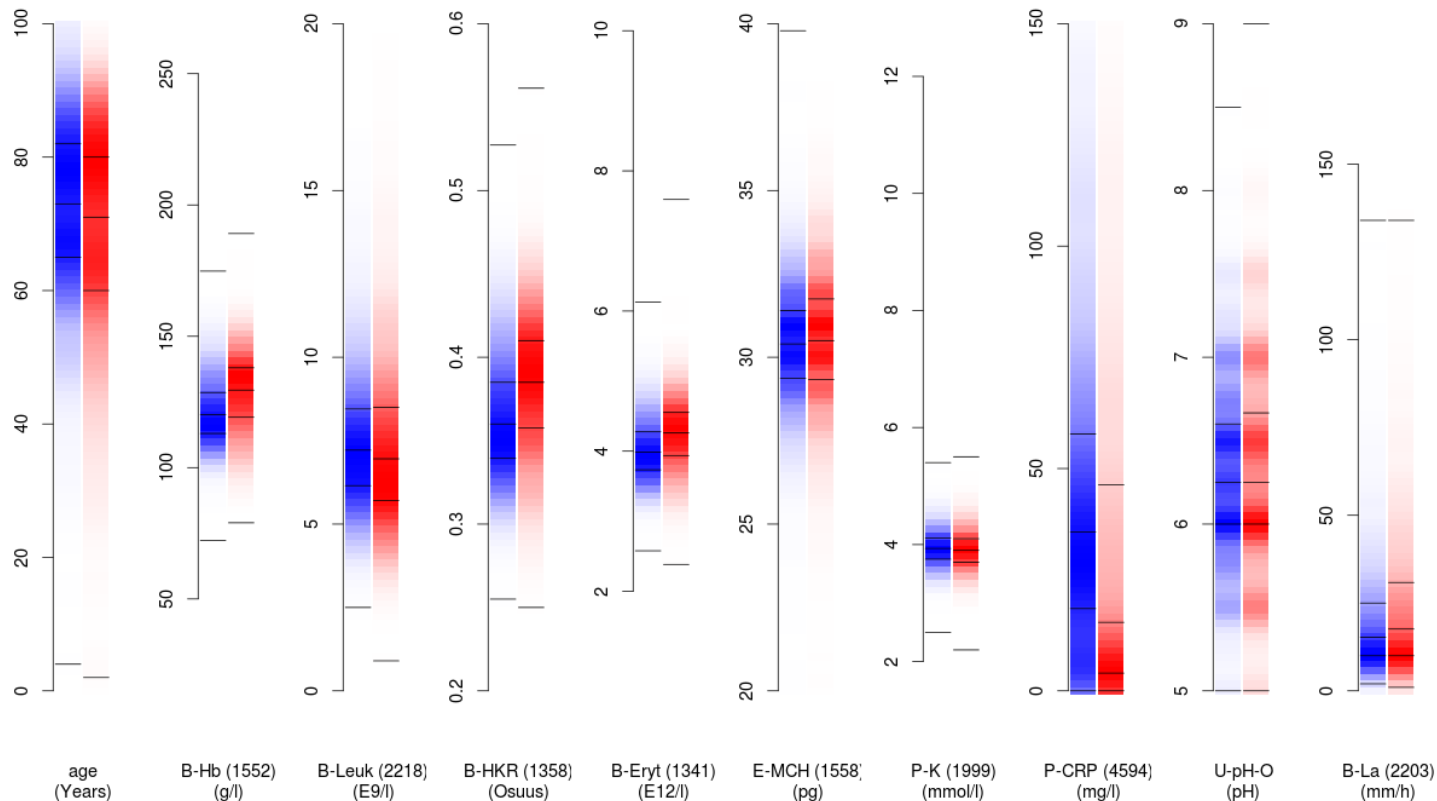
Some laboratory data seems to be missing.... There must be an error in the data transformation process.



Implementing Data Visualizations

Implementation can be divided into three levels

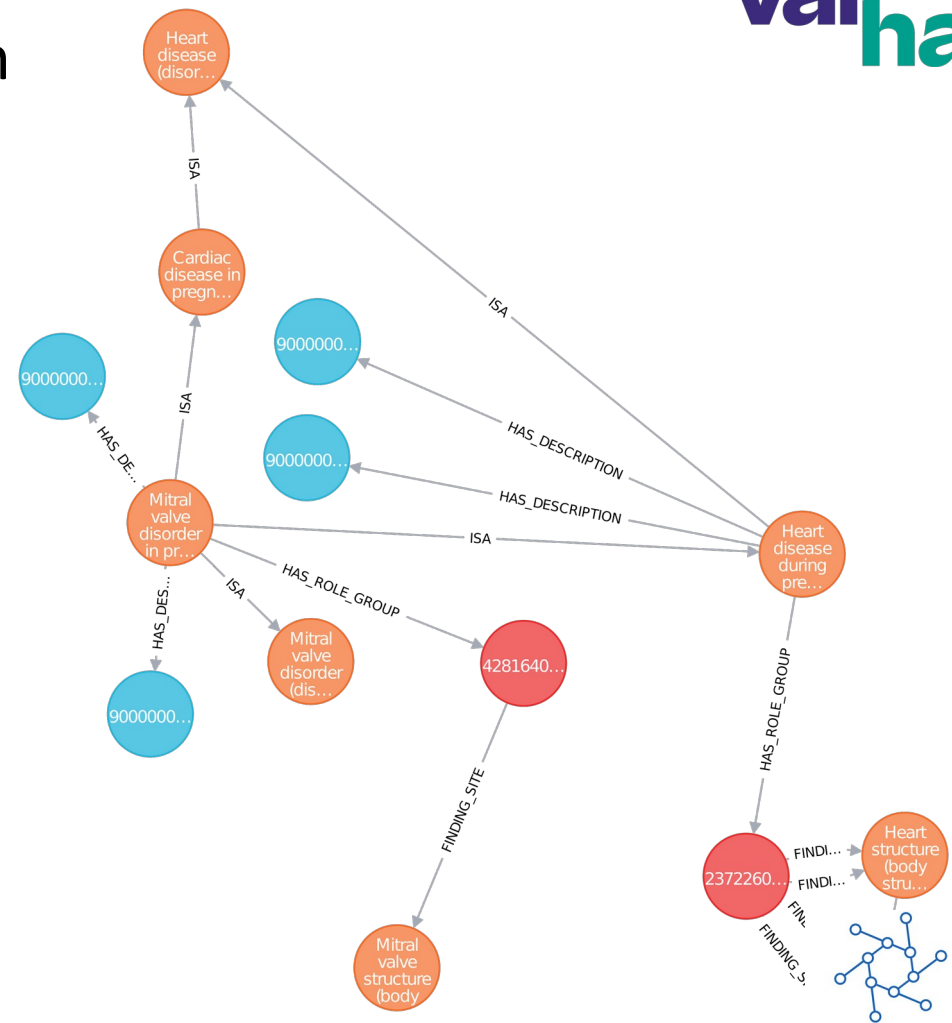
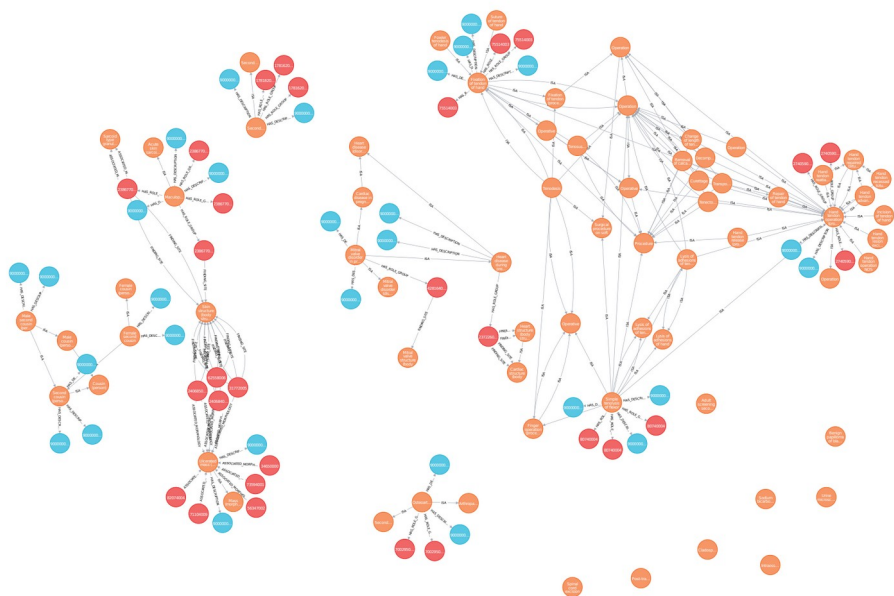
- 1) Self-service BI (PowerBI, Superset, Knowage, Qlik, Tableau,...)
- 2) Scripting with ready-made packages (R, Python,...)
- 3) Graphics programming (D3.js, R graphics primitives; analytical geometry)



Example: Visualization of the difference between blood values (R graphics)

More Examples of Data Visualization

Graphs can visualize dependencies between entities



Example: Snomed-CT code visualization with Neo4j-graph tool

Fundamentals of Statistical Visualizations

For recap – these are the tools of a visualist

length



area



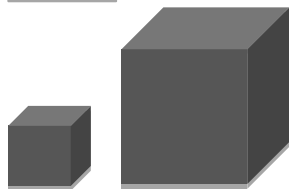
angle



shape



volume



color



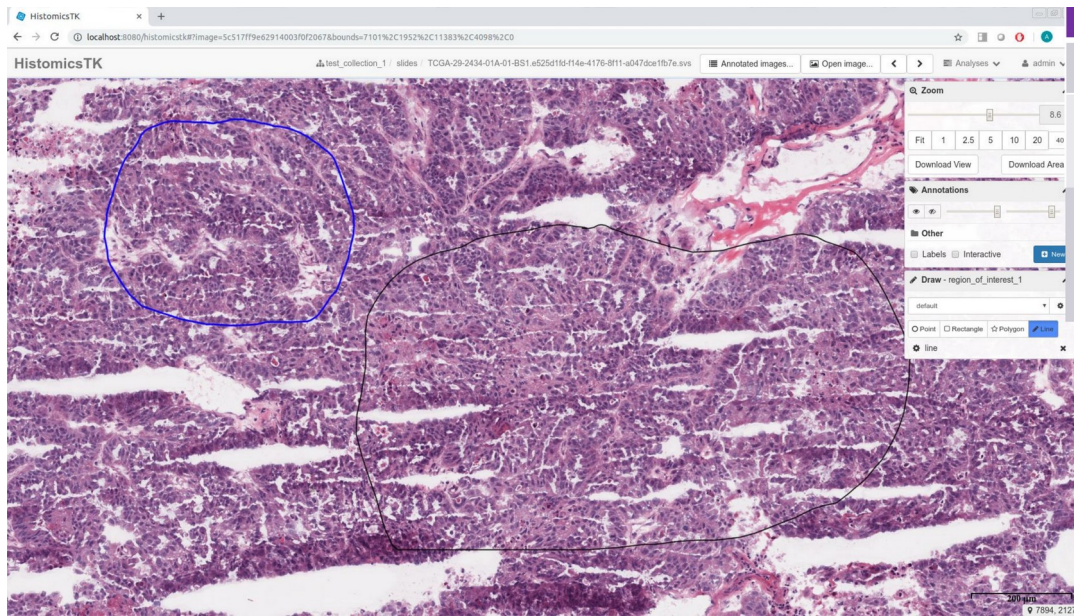
curvature



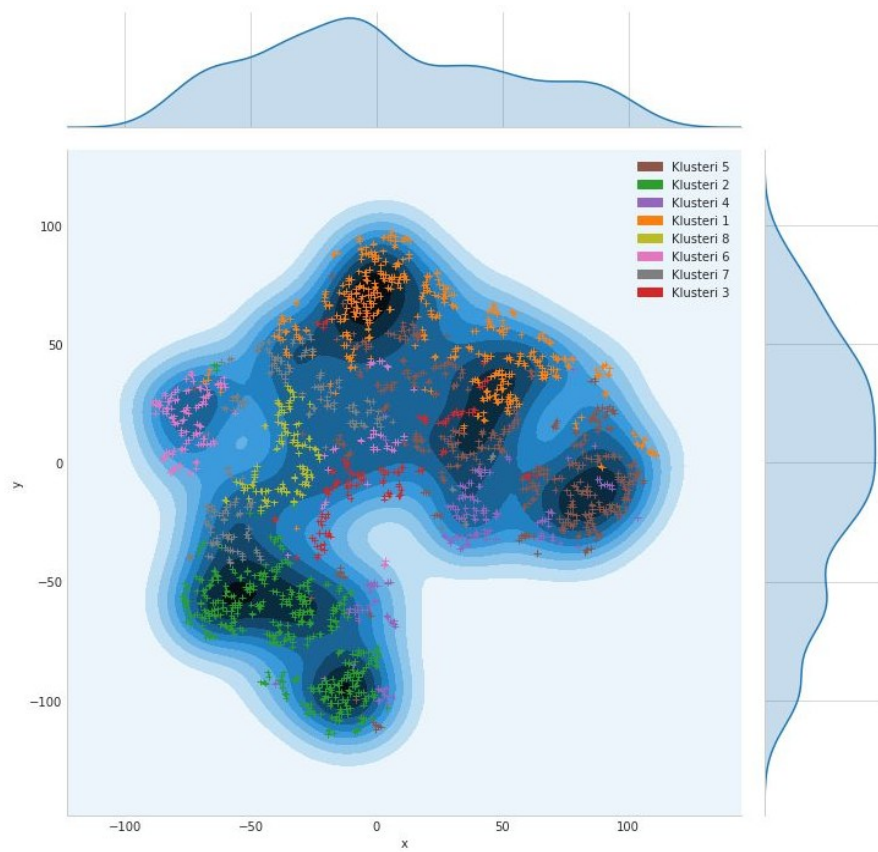
position



Predictive Analytics with Machine Learning / AI



Classification of pathological images in Auria Biobank



Example: Clustering of the different phenotypes of asthma in Auria Biobank

RESEARCH

Open Access

Automated detection of pulmonary embolism from CT-angiograms using deep learning

Heidi Huhtanen^{1*}, Mikko Nyman¹, Tarek Mohsen², Arho Virkki^{3,4}, Antti Karlsson⁵ and Jussi Hirvonen¹

Example

Publication available at:
<https://doi.org/10.1186/s12880-022-00763-z>

Abstract

Background: The aim of this study was to develop and evaluate a deep neural network model in the automated detection of pulmonary embolism (PE) from computed tomography pulmonary angiograms (CTPAs) using only weakly labelled training data.

Methods: We developed a deep neural network model consisting of two parts: a convolutional neural network architecture called InceptionResNet V2 and a long-short term memory network to process whole CTPA stacks as sequences of slices. Two versions of the model were created using either chest X-rays (Model A) or natural images (Model B) as pre-training data. We retrospectively collected 600 CTPAs to use in training and validation and 200 CTPAs to use in testing. CTPAs were annotated only with binary labels on both stack- and slice-based levels. Performance of the models was evaluated with ROC and precision–recall curves, specificity, sensitivity, accuracy, as well as positive and negative predictive values.

Results: Both models performed well on both stack- and slice-based levels. On the stack-based level, Model A reached specificity and sensitivity of 93.5% and 86.6%, respectively, outperforming Model B slightly (specificity 90.7% and sensitivity 83.5%). However, the difference between their ROC AUC scores was not statistically significant (0.94 vs 0.91, $p = 0.07$).

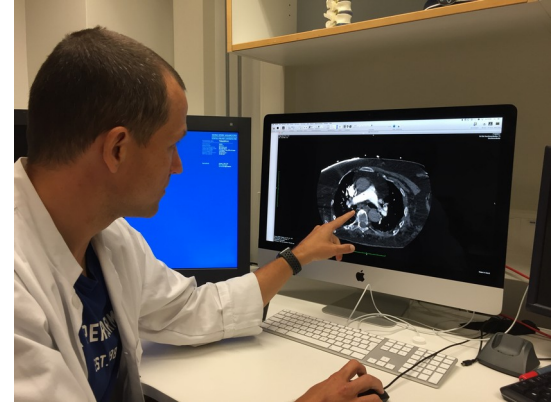
Conclusions: We show that a deep learning model trained with a relatively small, weakly annotated dataset can achieve excellent performance results in detecting PE from CTPAs.

Keywords: Artificial intelligence, Emergency radiology, Pulmonary embolism, Deep learning, Automated detection



Automated detection of pulmonary embolism

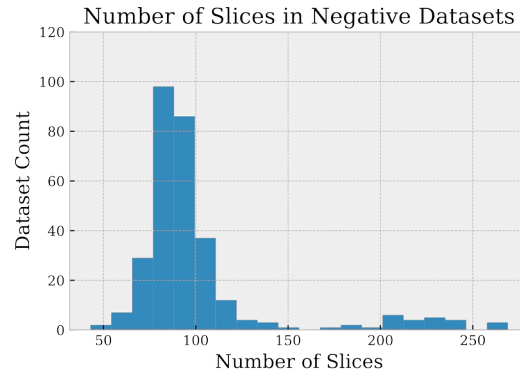
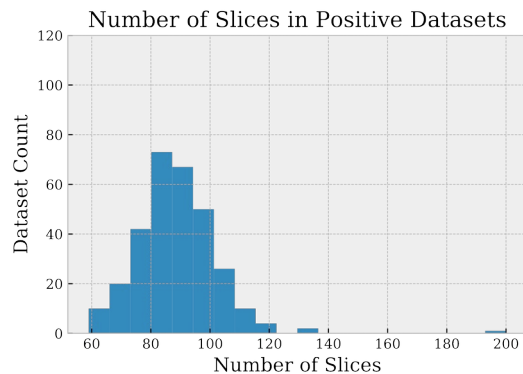
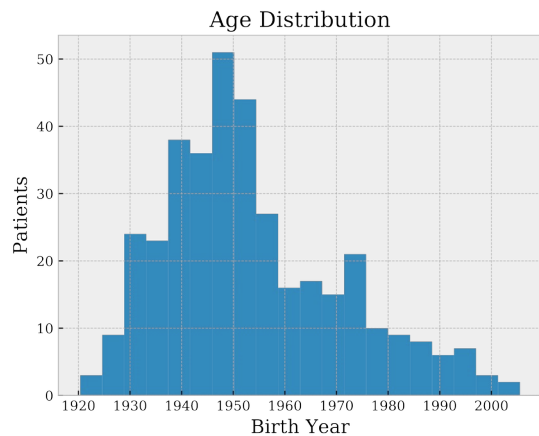
Motivation: If the patient can be diagnosed and treated < 10 minutes after arrival in the ICU, the probability of surviving increases significantly.



Dataset (58 057 slices)

2.1 Dataset

In this work a balanced dataset of 610 contrast enhanced CT images from 569 patients was collected retrospectively from period between October 2016 and September 2018. As the numbers imply, some patients had multiple images in the dataset due to re-takes or follow-ups of a chronic condition. Although most of the original DICOM studies were imaged using 1 millimeter slices with projections in axial, sagittal and coronal directions, we decided to restrict the scope to 3 millimeter reformatted axial slices. The purpose for this was to keep the amount of needed annotation work reasonable. On average each dataset contained of approximately 90 reformatted axial slices. In total there were 58057 individual slices from which 7229 had positive findings. The dataset contained images from 319 females and 250 males.



Model Technical Structure

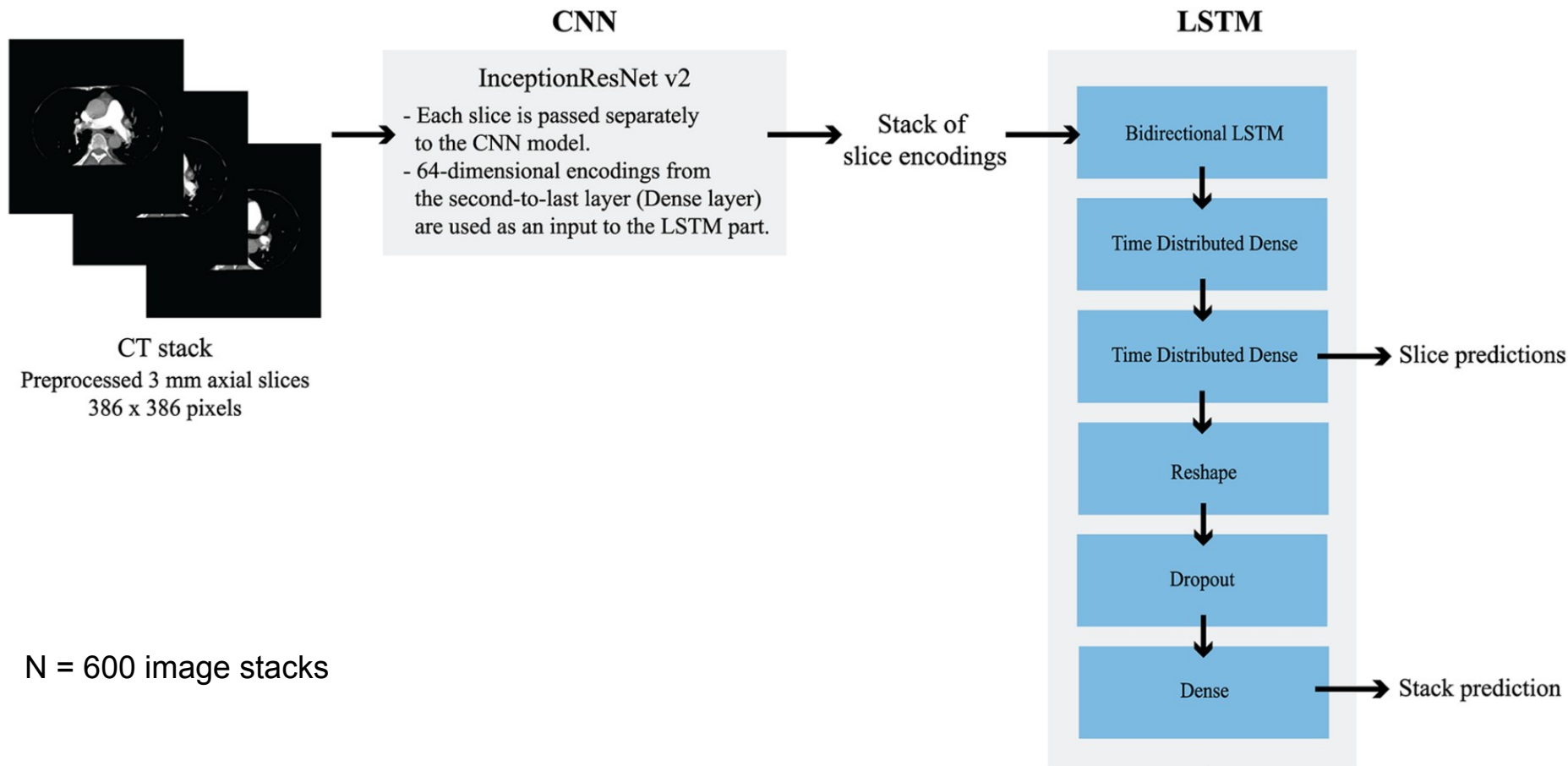


Fig. 1 Scheme of the model architecture

Results

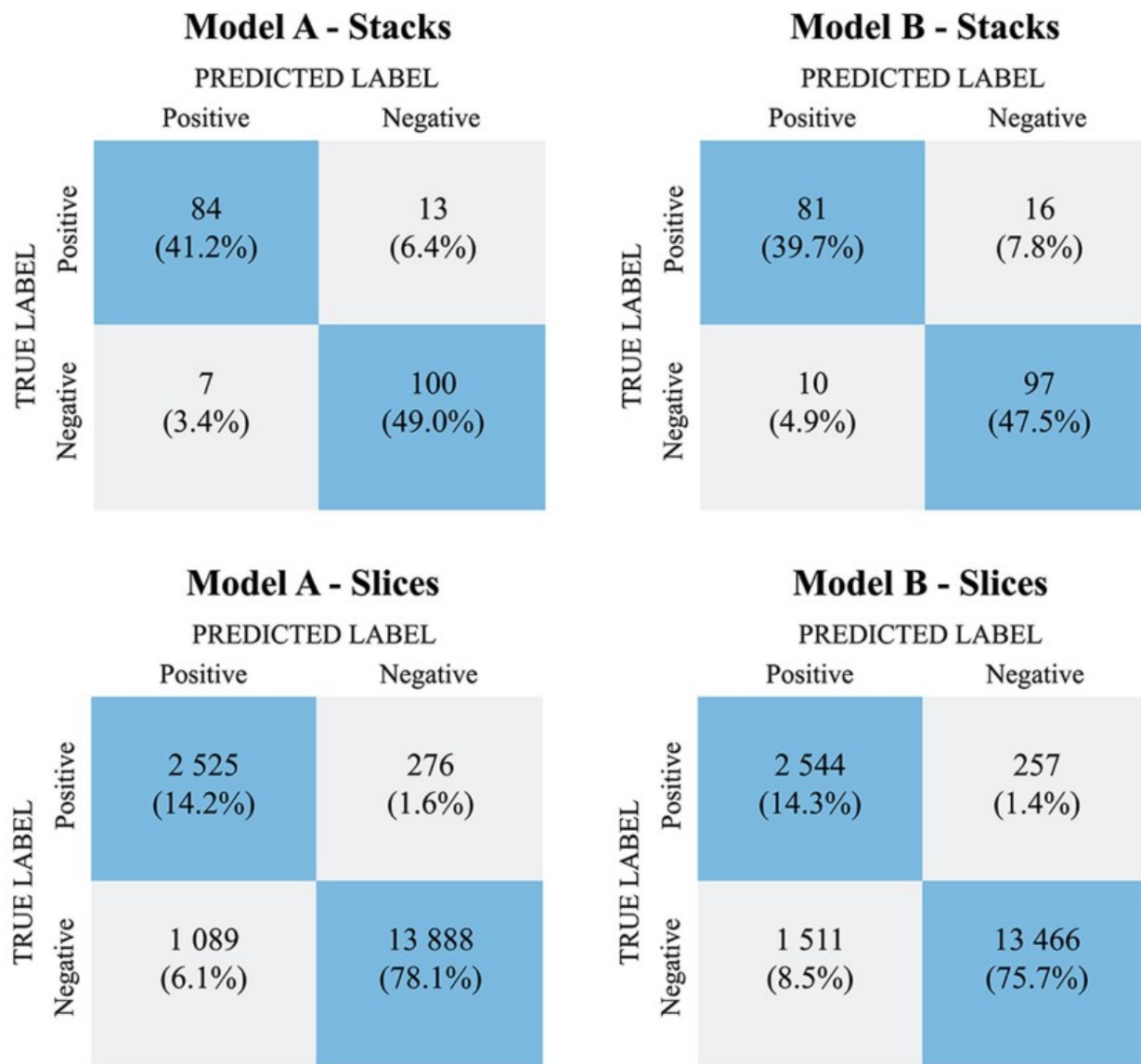
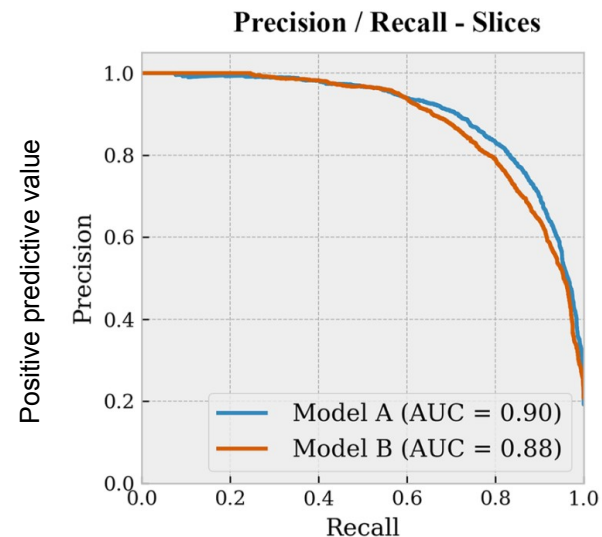
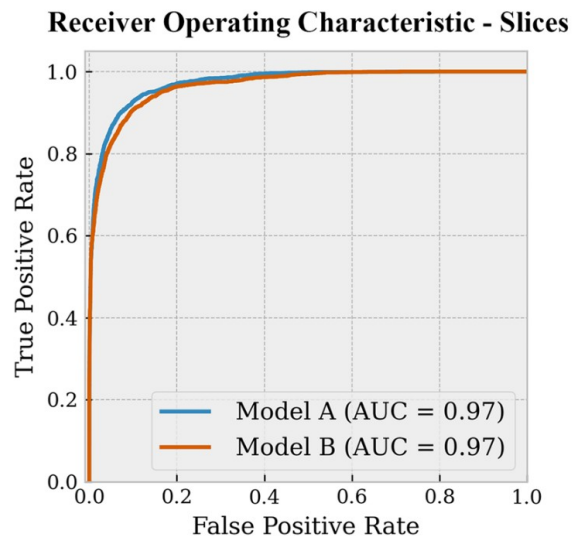
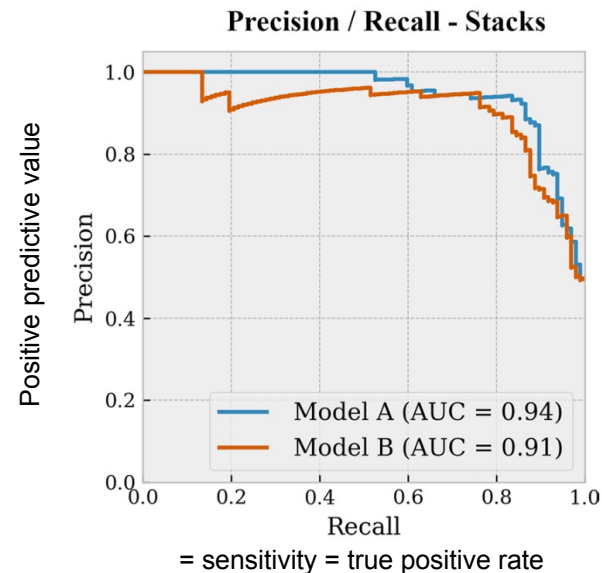
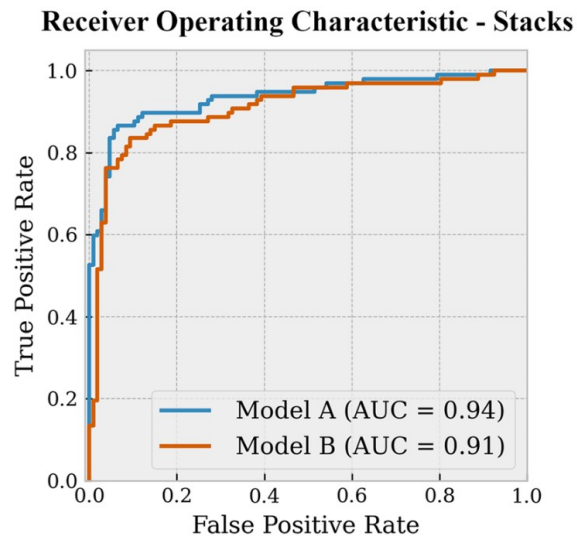


Fig. 3 Confusion matrices for Models A and B on stack- and slice-based classification

Results



Increased sensitivity
drops the positive
predictive value

Good models yield
high sensitivity and
positive predictive
value at the same
time

Fig. 2 ROC and PR curves for stack-based and slice-based predictions

Results

Discussion

The main contribution of this study is to demonstrate the development of a deep learning model for automated detection of PE from CTPAs, and to show that this is feasible even with limited data annotation resources. We found that our best model (Model A) achieved an ROC AUC of 0.94, sensitivity of 86.6% and specificity of 93.5% in predicting PE from whole CTPA stacks. We showed that these promising results could be achieved using a weakly labelled training dataset consisting of only 600 CTPAs, which is a relatively small dataset for neural networks.

Full article available at
<https://doi.org/10.1186/s12880-022-00763-z>



Hot Topics in 2025, Second Half

Application of open source'd in-house LLM's for

- Automated detection and reporting of adverse events
- Automated computation of quality metrics, e.g. clavien-dindo classification of surgical complications
- Confidential discussions with LLM's with augmented organizational or patient data
- ...

State of art LLM's could not be run in house in December 2024.

But now, in July 2025, Local FOSS LLM's compete with cloud computing.

```
arho@studio-nlp:~$ nvidia-smi
Tue Jul 1 15:54:39 2025
+-----+
| NVIDIA-SMI 550.163.01                  Driver Version: 550.163.01      CUDA Version: 12.4     |
+-----+-----+-----+-----+-----+-----+
| GPU  Name            Persistence-M   Bus-Id        Disp.A    Volatile Uncorr. ECC |
| Fan  Temp            Perf             Pwr:Usage/Cap     Memory-Usage     GPU-Util  Compute M. |
|=====+=====+=====+=====+=====+=====+
| 0   NVIDIA A100-SXM4-80GB      Off          00000000:07:00:0 Off      0             0%      Default |
| N/A   29C             P0              72W / 500W           17MiB / 81920MiB             Disabled |
+-----+-----+-----+-----+-----+-----+
| 1   NVIDIA A100-SXM4-80GB      Off          00000000:08:00:0 Off      0             0%      Default |
| N/A   29C             P0              78W / 500W           17MiB / 81920MiB             Disabled |
+-----+-----+-----+-----+-----+-----+

```

```
arho@studio-nlp:~$ ollama list
NAME                                ID                                SIZE    MODIFIED
gemma3:12b                          f4031aab637d                     8.1 GB  4 weeks ago
llama4:latest                       4f01ed6b6e01                     67 GB   5 weeks ago
alibayram/medgemma:latest          970df2db7db7                     2.5 GB  5 weeks ago
qwen3:0.6b                          3bae9c93586b                     522 MB  7 weeks ago
gemma3:4b                           a2af6cc3eb7f                     3.3 GB  7 weeks ago
gemma3:1b                           8648f39daa8f                     815 MB  7 weeks ago
llama4:scout                       4f01ed6b6e01                     67 GB   7 weeks ago
mistral:latest                     f974a74358d6                     4.1 GB  3 months ago
gemma3:27b                          a418f5838eaf                     17 GB   3 months ago
arho@studio-nlp:~$
```



Discussion





Auria Clinical
Informatics
<https://auria.fi/en>



AURIA CLINICAL
INFORMATICS