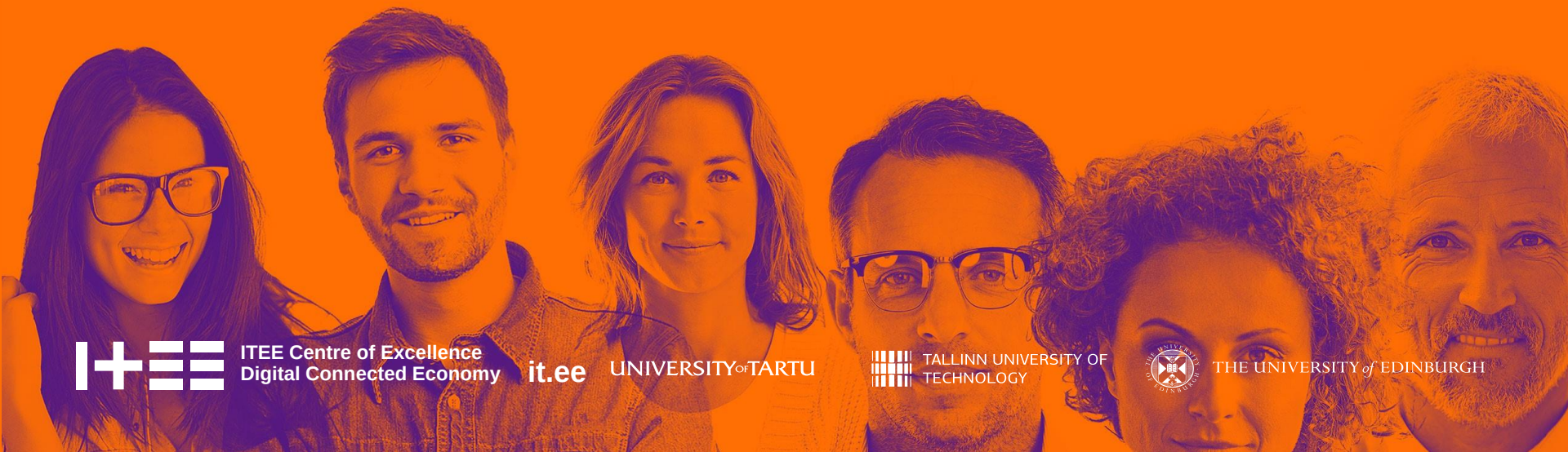
UNIVERSITY OF TARTU
1632

(Big) Health Data for Research

Jaak Vilo



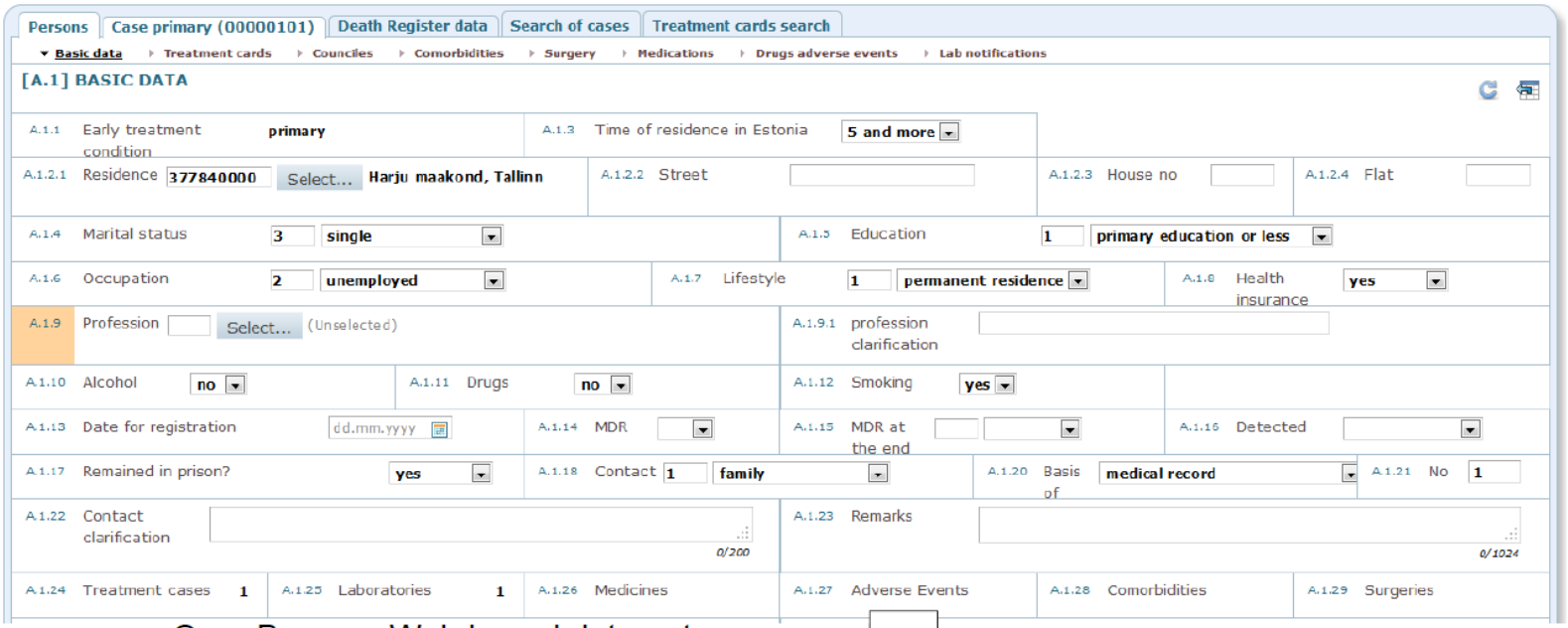
Founder and a PI of three teams

bioinformatics & data mining research 2002

Software for data management 2002

University-Business collaboartive Competence Center (CC)

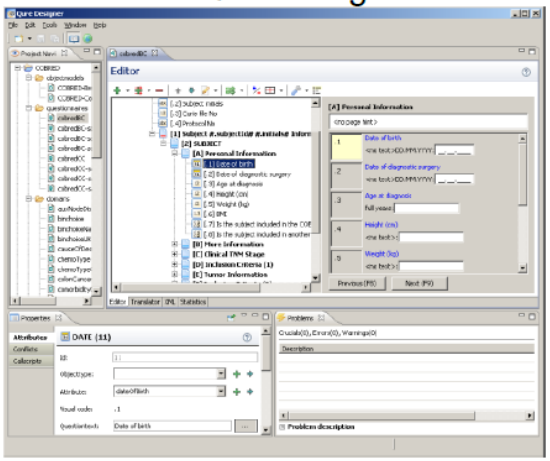
2009



Qure Browser Web based data entry

Load study definition XML files

Qure Designer



Qure Server

Relational database

Qure DataView
Ad-hoc queries
data output

Qure Analytics
Basic statistics
OLAP

Qure Reports
Predefined
reports

Estonian Health Registries on Qure Data Management Platform

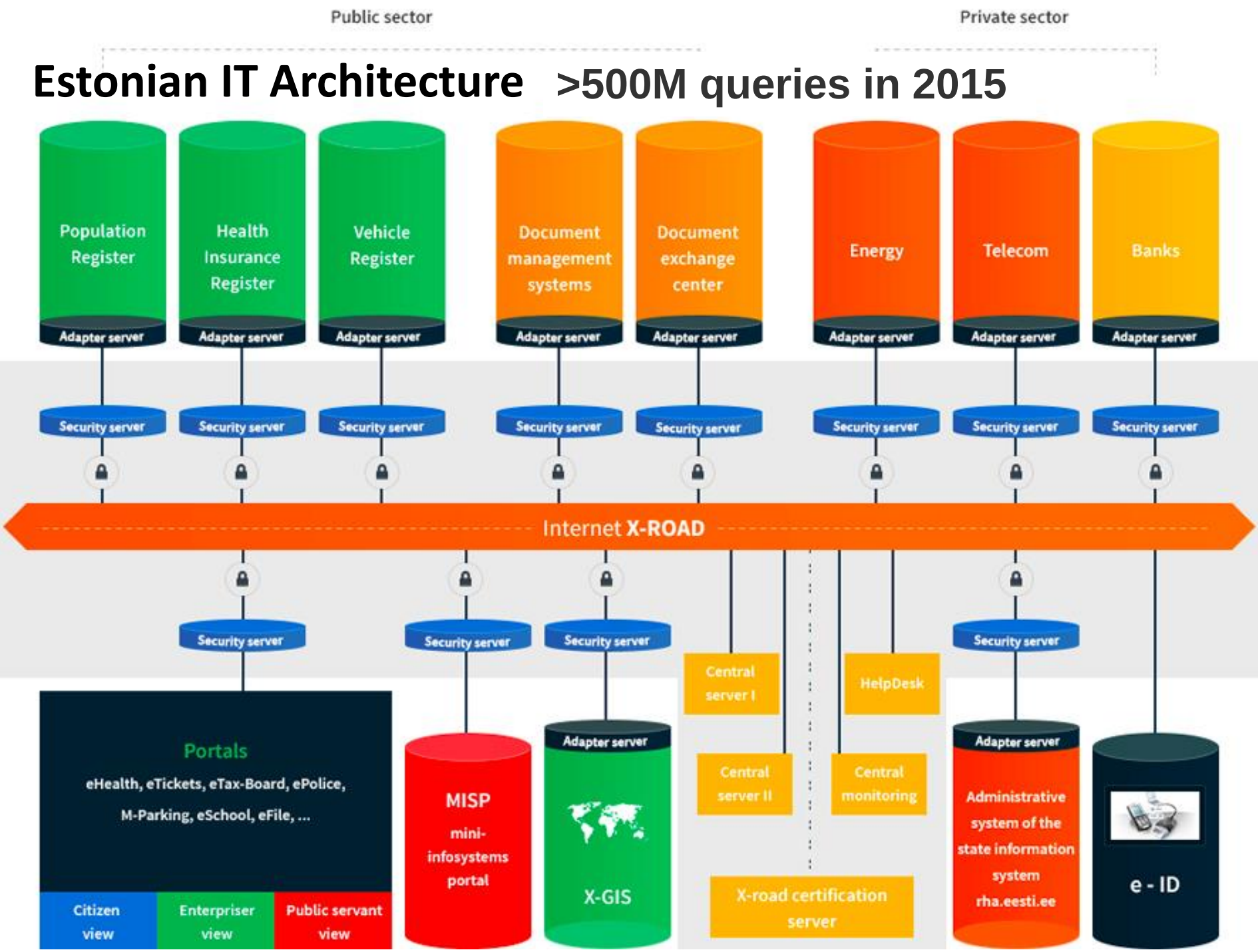
- Estonian Genome Project (biobank)
- Cancer registry
- Tuberculosis registry
- Medical birth registry
- Abortion registry
- Causes of death registry
- Drug treatment database
- HIV registry

National census 2011
67% online

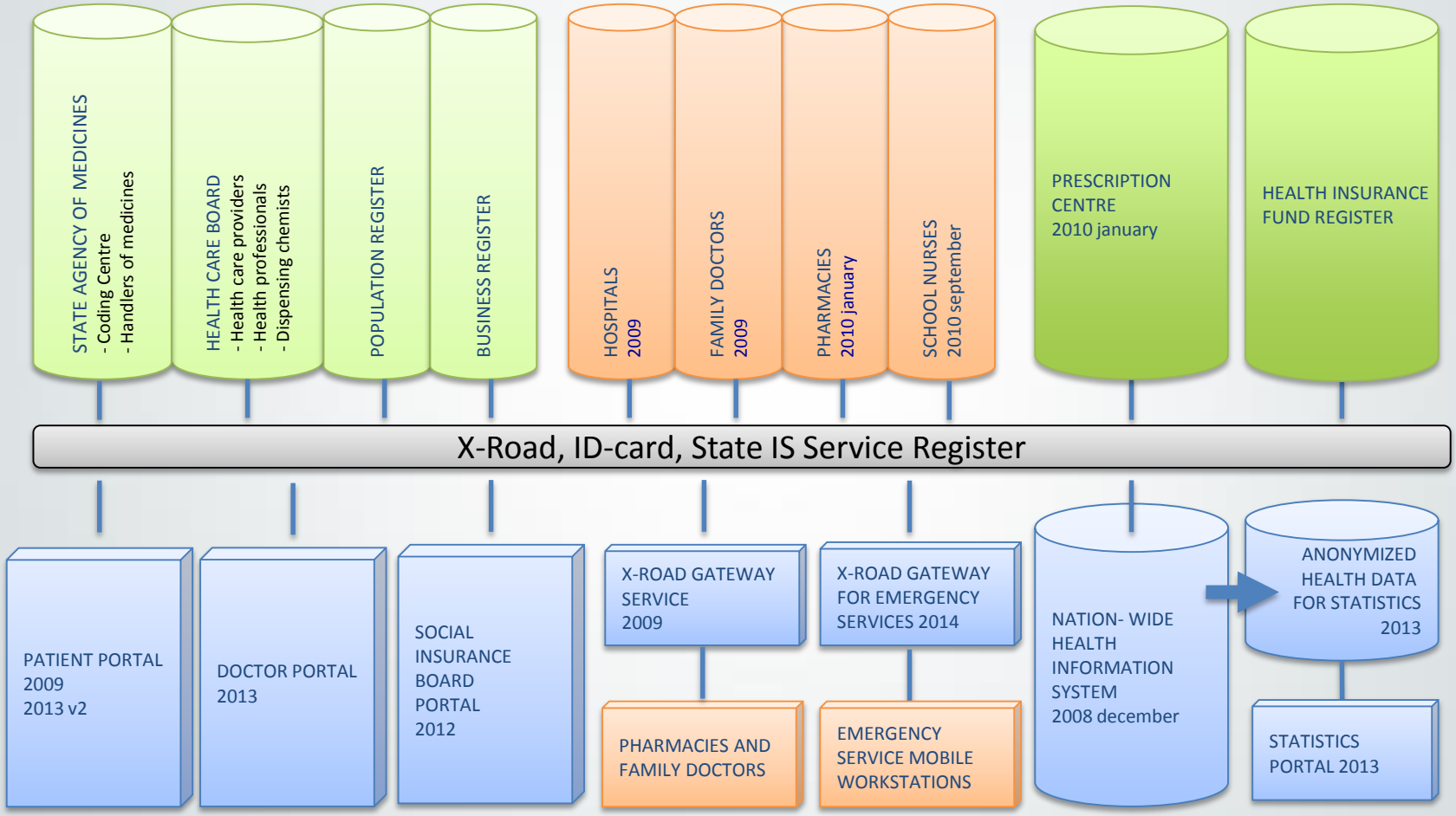
QureTEC

Estronian ID-code, e-ID, mobile ID

Estonian IT Architecture >500M queries in 2015



Estonian eHealth architecture



HL7v3 XML example 1

```
...
        <title>Lõplik kliiniline diagnoos</title>
        <!-- tekstilise kirjelduse algus-->
        <text>
          <paragraph>
            <content>1 I10 Arteriaalne hüpertoonia (Hüpertooniatõbi e essentsiaalne e primaarne arteriaalne hüpertension)
4 M15.8 Polüartroos (Muud polüartroosid)
4 K21.0 Ösofagiidiga reflukshaigus (ösofagiit gastro-ösofageaalse tagasivooluhaigusega)
4 I87.2 Krooniline veenipuudulikkus ((Perifeerne; krooniline) veenipuudulikkus)
</content>
          </paragraph>
        </text>
        <!-- tekstilise kirjelduse lõpp -->
        <entry typeCode="COMP" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
          <observation classCode="OBS" moodCode="EVN">
            <!-- eristab mis liiki Observation kirjega on tegemist (tehniline)-->
            <code code="DGN" codeSystem="1.3.6.1.4.1.28284.6.2.2.5.1" codeSystemName="Observation liik" displayName="Diagnoos" />
            <!-- staatus: lõplik-->
            <statusCode code="completed" />
            <!-- diagnoosi RHK-10 kood ja diagnoosi nimetus RHK-10 järgi -->
            <value code="I10" codeSystem="1.3.6.1.4.1.28284.6.2.1.13.1" codeSystemName="RHK-10" displayName="Hüpertooniatõbi e
essentsiaalne e primaarne arteriaalne hüpertension" xsi:type="CD">
              <!-- arsti sõnaline diagnoos -->
              <originalText>Arteriaalne hüpertoonia</originalText>
              <!-- diagnoosi statistiline liik  -->
              <qualifier>
                <value code="-" codeSystem="1.3.6.1.4.1.28284.6.2.1.1.1" codeSystemName="Diagnoosi statistiline liik"
displayName="kordusjuht elus" />
              </qualifier>
              <!-- diagnoosi statistiline liik lõpp-->
            </value>
            <!-- diagnoosi liik (põhihaigus)-->
            <interpretationCode code="MAIN" codeSystem="1.3.6.1.4.1.28284.6.2.1.2.1" codeSystemName="Diagnoosi liik"
displayName="Põhihaigus" />
          </observation>
        </entry>
        <entry typeCode="COMP" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
          <observation classCode="OBS" moodCode="EVN">
            <!-- eristab mis liiki Observation kirjega on tegemist (tehniline)-->
            <code code="DGN" codeSystem="1.3.6.1.4.1.28284.6.2.2.5.1" codeSystemName="Observation liik" displayName="Diag
            <!-- staatus: lõplik-->
            <statusCode code="completed" />
            <!-- diagnoosi RHK-10 kood ja diagnoosi nimetus RHK-10 järgi -->
...
```

DGN=Tegemist on diagnoosiga

Diagnoosi kood (RHK-10 klassifikaator)

Arsti sõnaline diagnoos (ei pruugi kattuda RHK-10 koodi ametliku nimetusega)

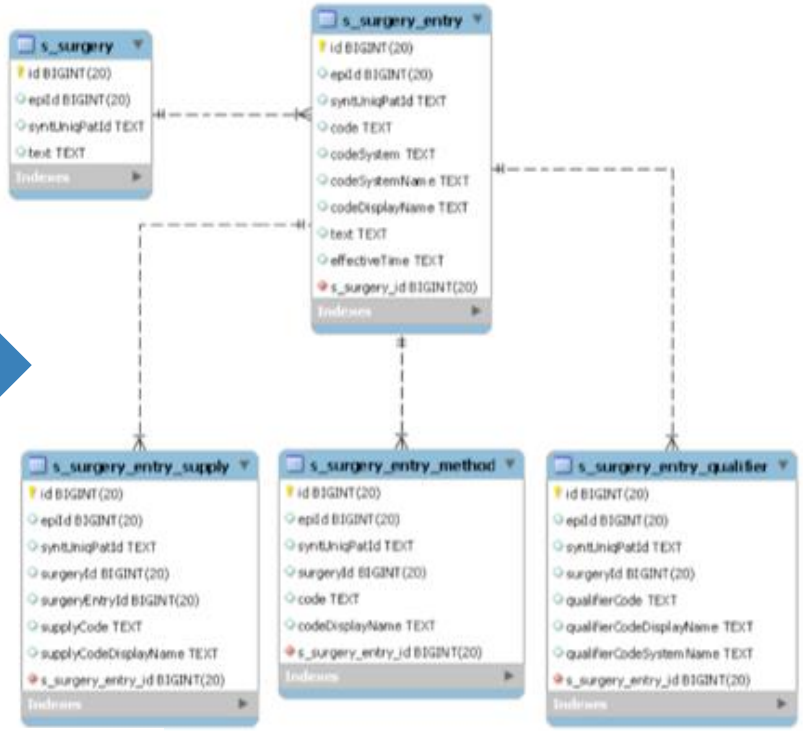
Diagnoosi statistiline liik

Tegemist on põhidiagnoosiga

XML -> DB

Operatsioonid

Kuupäev	Operatsiooni koodid NCSP	Alternatiivkood	Anesteesia liik	Lisavahendid
11.02.2007	2235 - Totaalne endoproteesimine põlveliigesel 1129 - Põlveliigese esmane proteesimine totaalse proteesiga		endotrahheaalne anesteesia	2950L - puusaliigese endoprotees 2614L - reieluukaela osteosünteesi CSS-komplekt (3 kruvi)
Põlveliigese esmane proteesimine totaalse proteesiga kasutades tsementi Spinaal+epiduraalanesteesias haige seliliasndis parem põlveliiges avatud parapatellaarse ja 11,6 cm MMI lõikega. Pinnad saetud NexGen LPS E mudeli järgi ja tsementeeritud säärekomponent No5 ja reiekomponent E/R paigale. Lisaks säareplatoo 10 green. Haav loputatud, hemostaas. Dreen eraldi avast aspiratsiooniks. Haav kihiti suletud, nahale madratsõmblused. Aseptiline side. Elastne side.				
11.02.2007	DJD20 - Ninavaheseina plastiline korrektsioon ZXD10 - Plaaniline protseduur	041011 - Septoplastika (Haigekassa hinnakirjakoodid)	endotrahheaalne anesteesia	
Plaaniline ninaseina plastiline korrektsioon.				
11.02.2007	DHB50 - Konhoplastika ZXA00 - Parem pool ZXC60 - Kuumuse kasutamine	021008 - Submukoosne konhotoomia (Haigekassa hinnakirjakoodid)		



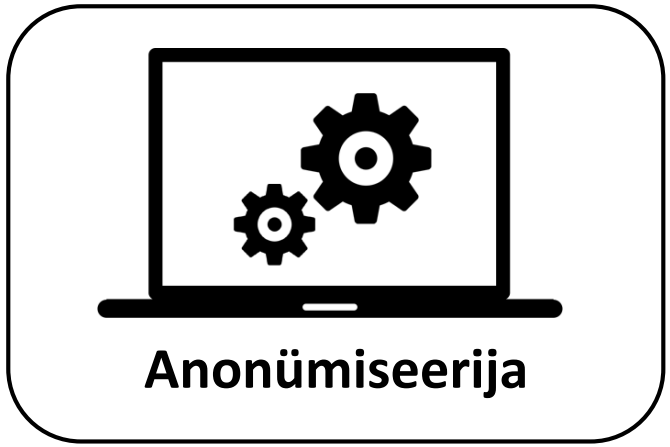

```
<!-- OPERATSIOONID ALGUS-->
<component typeCode="COMP">
        <section classCode="DOCSECT" moodCode="EVN">
                <!-- mis tüüpi sektsiooniga on tegu "SUR" -->
                <code code="SUR" codeSystem="1.3.6.1.4.1.28284.6.2.2.11.2"
codeSystemName="Sektsiooni kodeering" displayName="Operatsioonid"/>
                <!-- sektsiooni pealkiri -->
                <title>Operatsioonid</title>
                <!-- tekstilise kirjelduse algus-->
                <text>
                        <paragraph>
                                <content>
                                        Kuupäev: 11.02.2007
                                        Operatsiooni koodid: 2235 -
                                        Totaalne endoproteesimine põlveliigesel,
                                        1129 - Põlveliigese esmane proteesimine
                                        totaalse proteesiga
                                        Anesteesia liik : A01 - endotrahheaalne anesteesia
                                        isavahendid:
                                        2950L-puusaliigese endoprotees
                                        2614L-reieluukaela osteosünteesi
                                        CSS-komplekt (3 kruvi)
                                        Operatsiooni kirjeldus
                                         Põlveliigese
```

Tarkvara TAK (STACC)

Anonymisation from free text

Sisendtekst

Patsient **John Doe** Vanus 44 a. IK – **77771478888** võeti statsionaarsele ravile.
Asjaolude täpsustamiseks helistada dr. **Hämarikule** tel: **7177765**, kell 10.00-13.00.



95%
isikuinfost
eemaldatud

Anonümiseeritud tekst

Patsient **XXX** Vanus 44 a. IK – **XXX** võeti statsionaarsele ravile. Asjaolude täpsustamiseks helistada dr. **XXX** tel: **XXX**, kell 10.00-13.00.

What is meant by each abbreviation or acronym/abbreviation

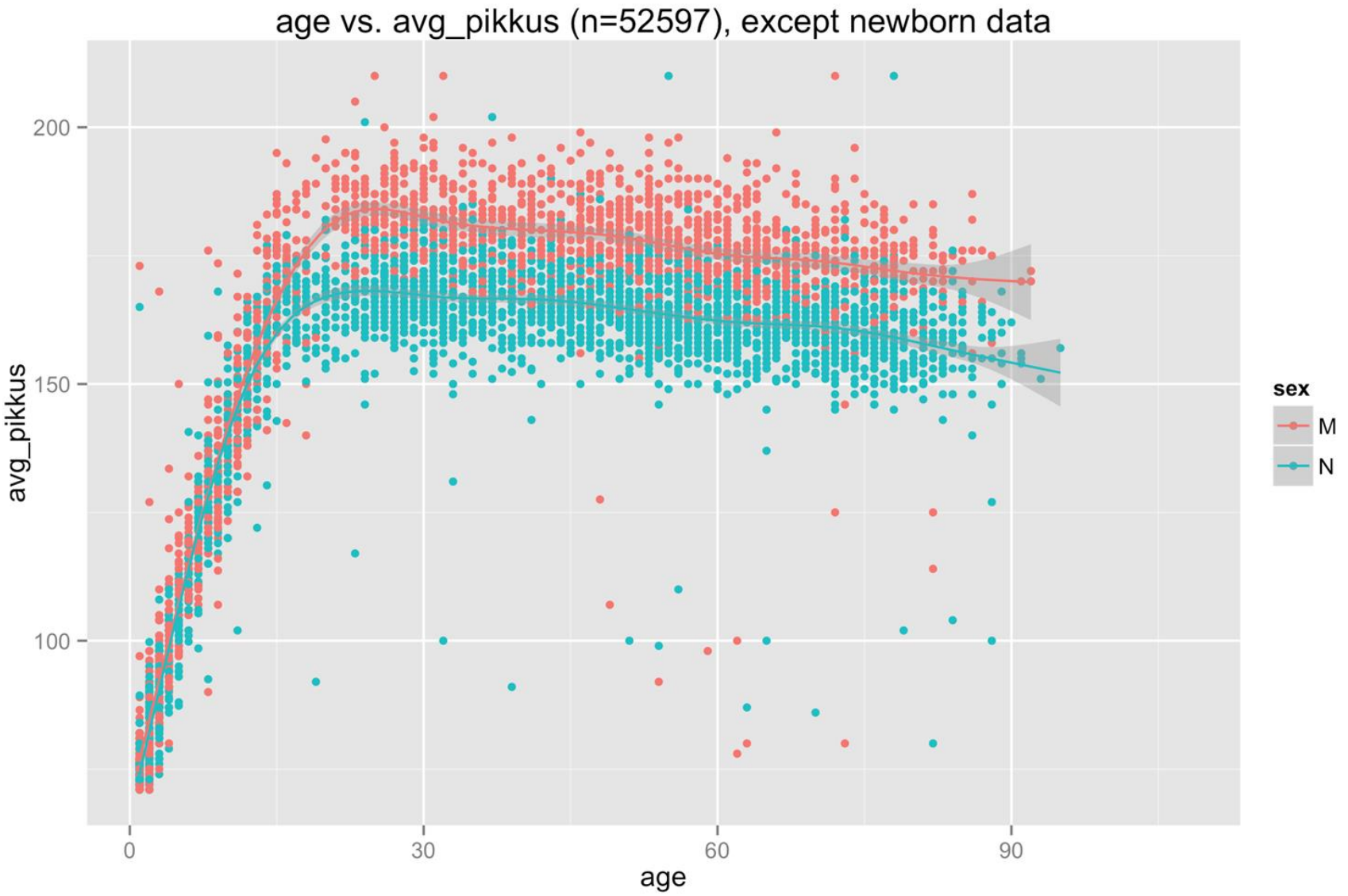
Kontekst	Täisvorm
… **p** silm ei näe …	**parem**
… kolmas **p** palavik …	**päev**
… vähene **p** pleurareaktsiooni riba …	**parietaalne**
… **p** 6mm ümargune …	**pupill**

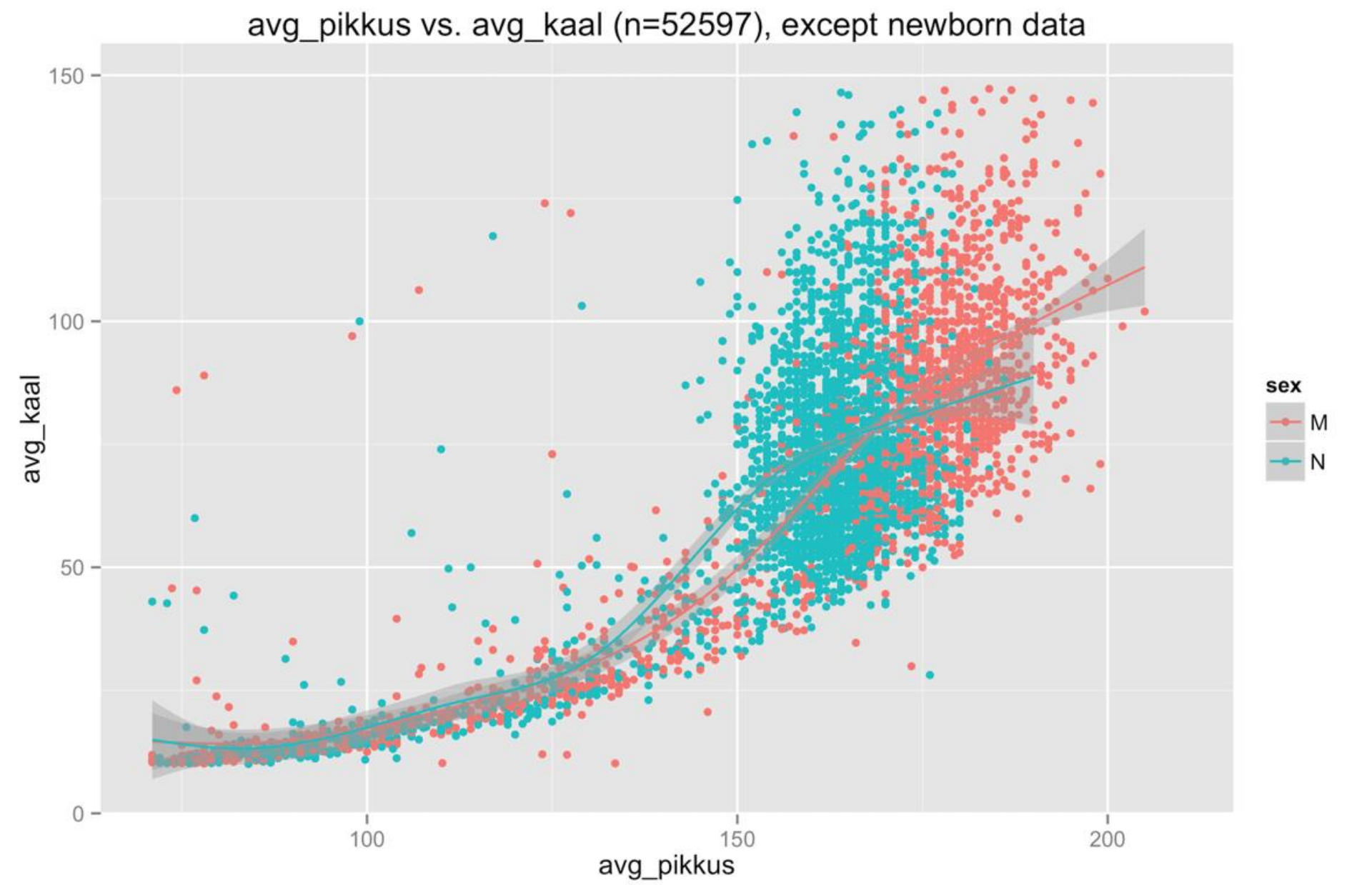
Täisvorm	Skoor
parietaalne	92%
parem	4%
päev	3%
pupill	0,3%
…	

… important to understand free text

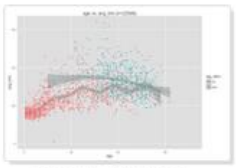
Free Text	**->**	**Information extraction**
•Siinusrütm **59** lööki min		59
•Kiirenenud, Ekg-l siinus rütm , **160**/min		160
•vatsakeste tahhüarütmia fr - ga **150** lööki min		150
•EKG --> siinusrütm, fr.**110**xmin		110
•Ps **66**/min		66
•Fr = **77** ' min		77
•Maksimaalne fr=**196** , minimaalne fr=**55**		**55 : 196**
•V/v-l **RR 133/105/90** , arütmia		**90**

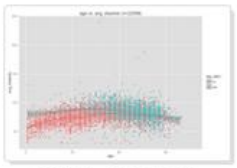

age vs. avg_pikkus (n=52597)
avg_pikkus
200
150
100
50
0
0
30
60
90
age
sex
M
N

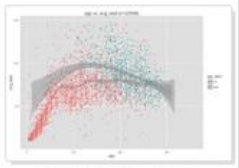

age vs. avg_pikkus (n=52597), except newborn data
avg_pikkus
200
150
100
0
30
60
90
age
sex
M
N

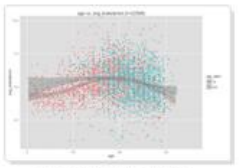

avg_pikkus vs. avg_kaal (n=52597), except newborn data
avg_kaal
150
100
50
0
100
150
200
avg_pikkus
sex
M
N

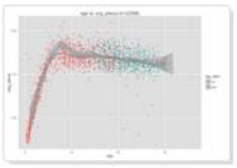
plots_i10i11 73 items

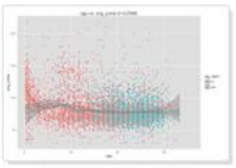

age vs. avg_bmi …


age vs. avg_diast…

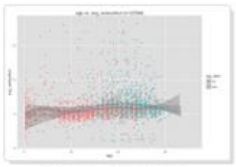

age vs. avg_kaal …

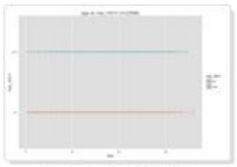

age vs. avg_kole…

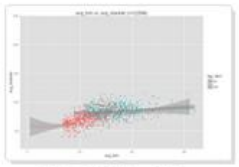

age vs. avg_pikk…

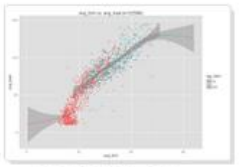

age vs. avg_puls…

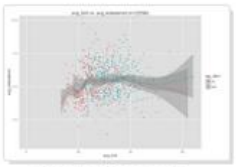

age vs. avg_syst…

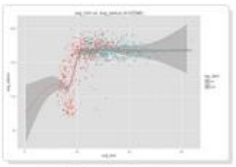

age vs. avg_vere…

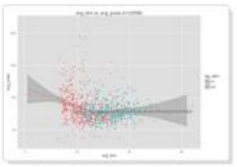

age vs. has_i10i1…

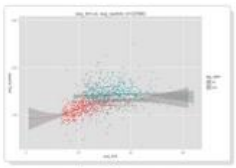

avg_bmi vs. avg_…


avg_bmi vs. avg_…

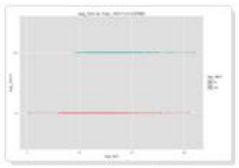

avg_bmi vs. avg_…

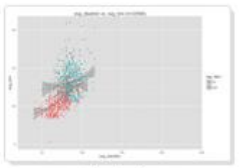

avg_bmi vs. avg_…

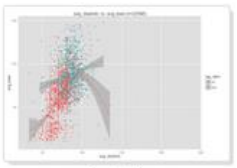

avg_bmi vs. avg_…

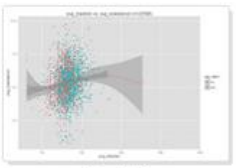

avg_bmi vs. avg_…

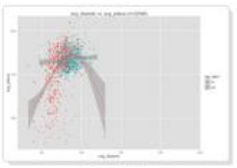

avg_bmi vs. avg_…


avg_bmi vs. has_…


avg_diastolic vs. …


avg_diastolic vs. …


avg_diastolic vs. …


avg_diastolic vs. …

Data: 23 million analyses entries

https://p12.stacc.ee/common/epicrisis/?id=123429792#S

Biokeemia analüüs	Rerentsväärtus	23.05.2012 09:15:00
Materjal		
fS-Gluc (glükoos)	4,1-5,9 mmol/L	5.8
S-K (kaalium)	3,5-5,1 mmol/L	4.5
S-VitD(25-OH) vitamiin D (25-OH) seerumis	>= 75 nmol/L	52.90
S-ALAT (alaniini aminotransferaas)	N: <31 U/L, M:<41 U/L	87
S-ASAT (aspartaadi aminotransferaas)	N:<32 U/L, M:<38 U/L	48
S-LDH (laktaadi dehüdrogenaas)	240-480 U/L	483
fS-Urea (uurea)	<=65a.<8,3 ; >65a <11,9 mmol/L	5.6
S-Bil (bilirubiin)	< 17,1 µmol/L	7
S-LDL-Chol (LDL kolesterool)	<3,0 mmol/L	4.71
S-Alb (albumiin)	35 - 52 g/L	43.2
S-GGT (gamma-glütamüüli transferaas)	N:<40 ; M:<60 U/L	59
S-Chol (kolesterool)	<5 mmol/L	6.2
S-HDL-Chol (HDL kolesterool)	>1,0 mmol/L	1.16
fS-Trigl (triglütseriidid)	<2 mmol/L	1.51
S-CRP (C-reaktiivne valk)	<5 mg/L	<1
S-UA (kusihape)	N:143-339,M:202-417 µmol/L	356
S-Na (naatrium)	136-145 mmol/L	143
S-Prot (valk)	66-87g/L	66.0
fS-Transf-sR (transferriini lahustuvad retseptorid)	N:1,9-4,4;M:2,2-5,0 mg/L	2.67
S-NT-proBNP (B-tüüpi natriureetilise propeptiidi N-fragment	<400pg/mL- kroonilise südamepuudulikkuse välistuspiir; >2000pg/mL- diagnost. otsustuspiir	96

Hemostasiogramm **Rerentsväärtus 23.05.2012 09:17:00**

Materjal

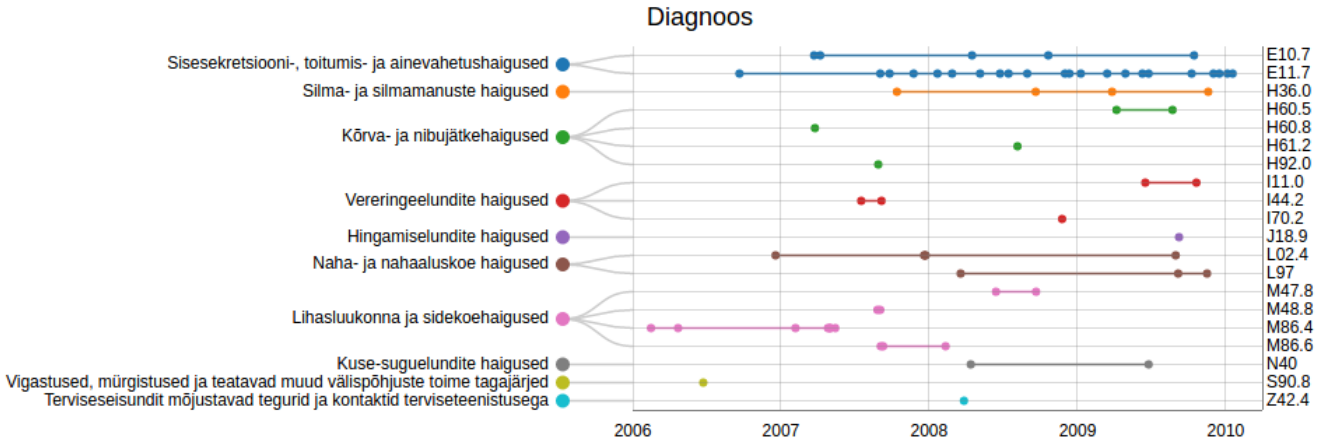
P-DDi - D-dimeerid <0,5 µg/ml 0.28

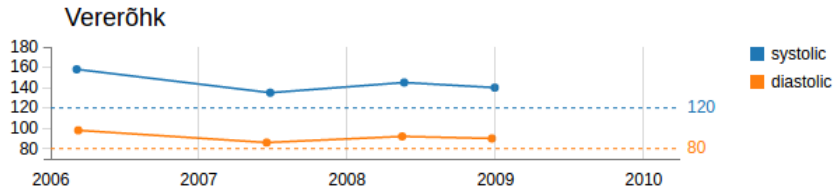
Immuunmeetoditel uuringud	Rerentsväärtus	23.05.2012 09:20:00
Materjal		seerum
TSH / Türeotropiin	0,4-4,0 mIU/L	2.36
FT3 / Vaba trijoodtüroniin	2,8-6,5 pmol/l	5.9
FT4 / Vaba türoksiin	11,5-22,5 pmol/L	12.4
ANA /Tuumavastane IgG seerumis	< 1:100	NEG
Tsüklilise tsitrulleeritud peptiidi vast IgG seerumis	<7,0 U/ml	1.0
IgE (Immuunoglobuliin E)	< 100 kU/l	

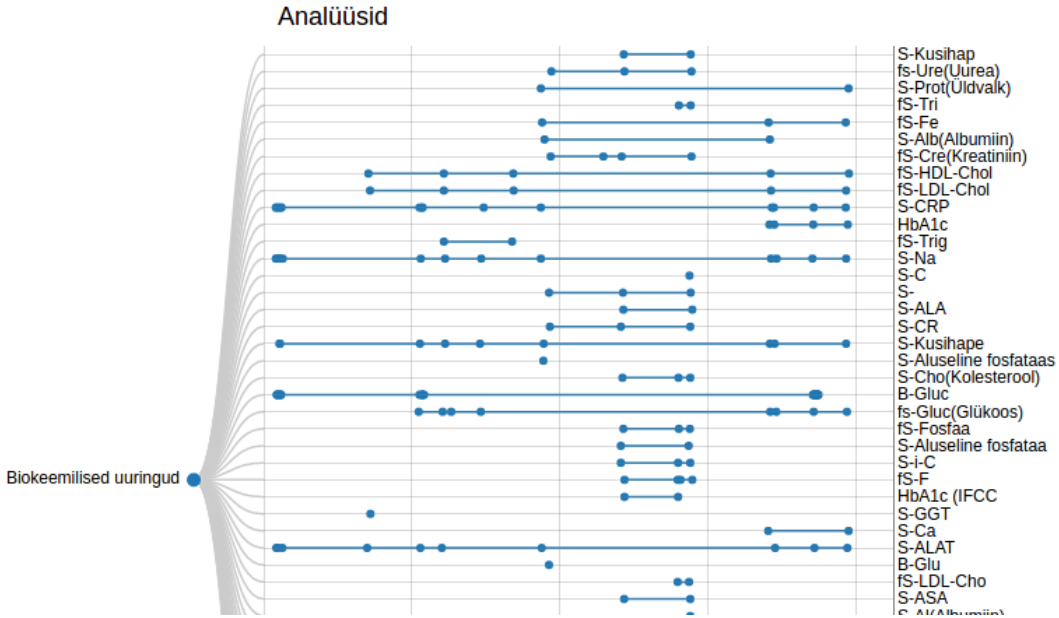
Uriini analüüs **Rerentsväärtus** **24.05.2012 07:00:00**

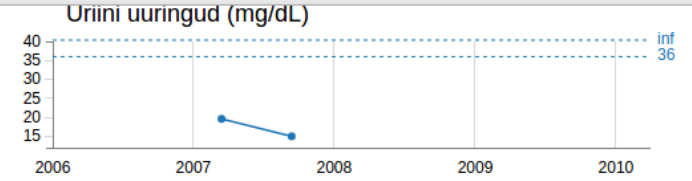
Materjal

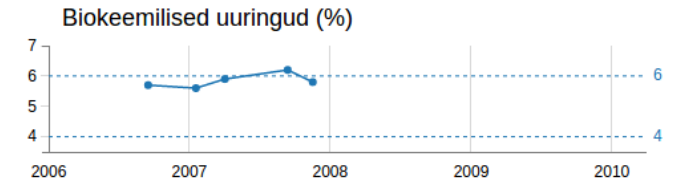
Uriini ribaanalüüs tehtud

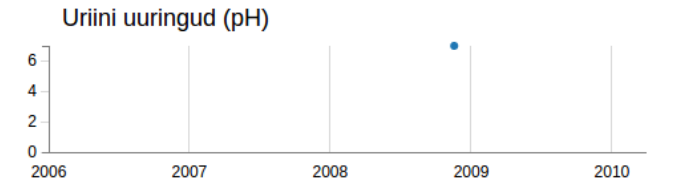

Diagnoos
Sisesekretsiooni-, toitumis- ja ainevahetushaigused
Silma- ja silmamanuste haigused
Kõrva- ja nibujätkehaigused
Vereringeelundite haigused
Hingamiselundite haigused
Naha- ja nahaaluskoe haigused
Lihasluukonna ja sidekoehaigused
Kuse-suguelundite haigused
Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega
E10.7
E11.7
H36.0
H60.5
H60.8
H61.2
H92.0
I11.0
I44.2
I70.2
J18.9
L02.4
L97
M47.8
M48.8
M86.4
M86.6
N40
S90.8
Z42.4
2006
2007
2008
2009
2010

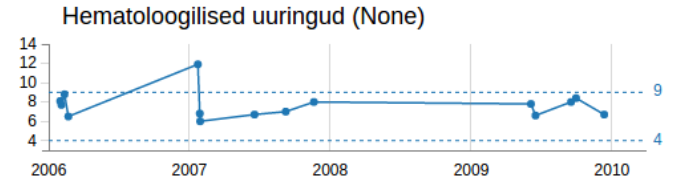

Vererõhk
180
160
140
120
100
80
120
80
systolic
diastolic
2006
2007
2008
2009
2010

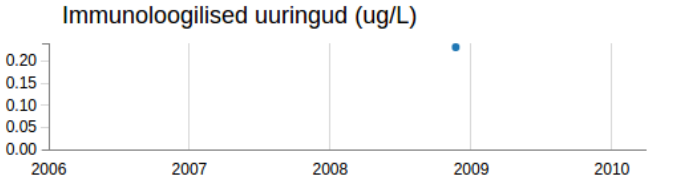

Analüüsid
Biokeemilised uuringud
S-Kusihap
fs-Ure(Uurea)
S-Prot(Üldvalk)
fS-Tri
fS-Fe
S-Alb(Albumiin)
fS-Cre(Kreatiniin)
fS-HDL-Chol
fS-LDL-Chol
S-CRP
HbA1c
fS-Trig
S-Na
S-C
S-
S-ALA
S-CR
S-Kusihape
S-Aluseline fosfataas
S-Cho(Kolesterool)
B-Gluc
fs-Gluc(Glükoos)
fS-Fosfaa
S-Aluseline fosfataa
S-i-C
fS-F
HbA1c (IFCC
S-GGT
S-Ca
S-ALAT
B-Glu
fS-LDL-Cho
S-ASA

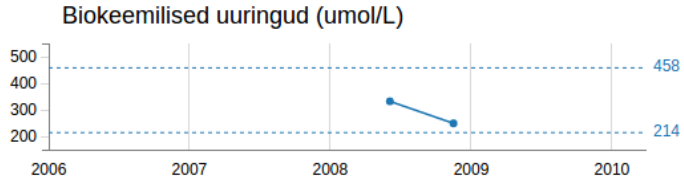

Uriini uuringud (mg/dL)
40
35
30
25
20
15
inf
36
2006
2007
2008
2009
2010

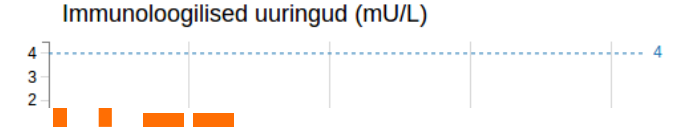

Biokeemilised uuringud (%)
7
6
5
4
6
4
2006
2007
2008
2009
2010

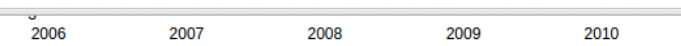

Uriini uuringud (pH)
6
4
2
0
2006
2007
2008
2009
2010

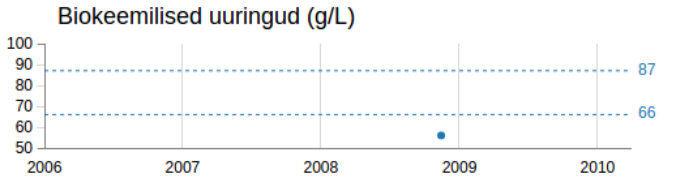

Hematoloogilised uuringud (None)
14
12
10
8
6
4
9
4
2006
2007
2008
2009
2010

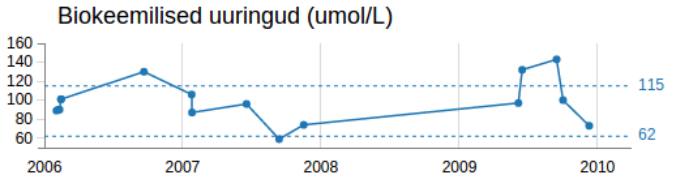

Immunoloogilised uuringud (ug/L)
0.20
0.15
0.10
0.05
0.00
2006
2007
2008
2009
2010

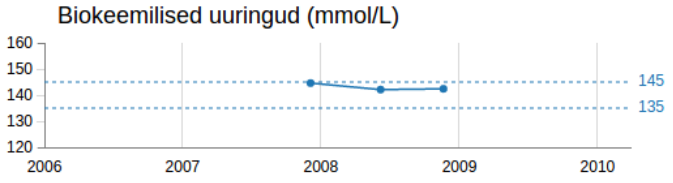

Biokeemilised uuringud (umol/L)
500
400
300
200
458
214
2006
2007
2008
2009
2010

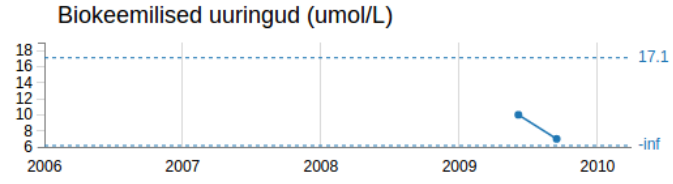

Immunoloogilised uuringud (mU/L)
4
3
2
4

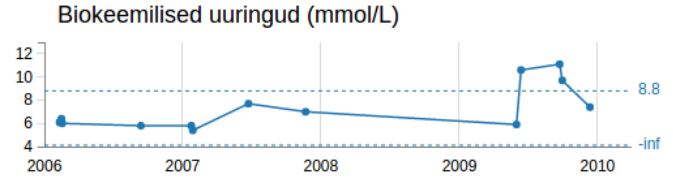

2006
2007
2008
2009
2010

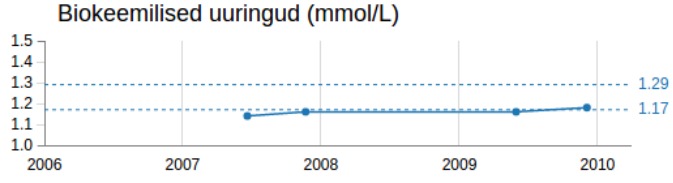

Biokeemilised uuringud (g/L)
100
90
80
70
60
50
87
66
2006
2007
2008
2009
2010

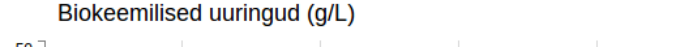

Biokeemilised uuringud (umol/L)
160
140
120
100
80
60
115
62
2006
2007
2008
2009
2010


Biokeemilised uuringud (mmol/L)
160
150
140
130
120
145
135
2006
2007
2008
2009
2010

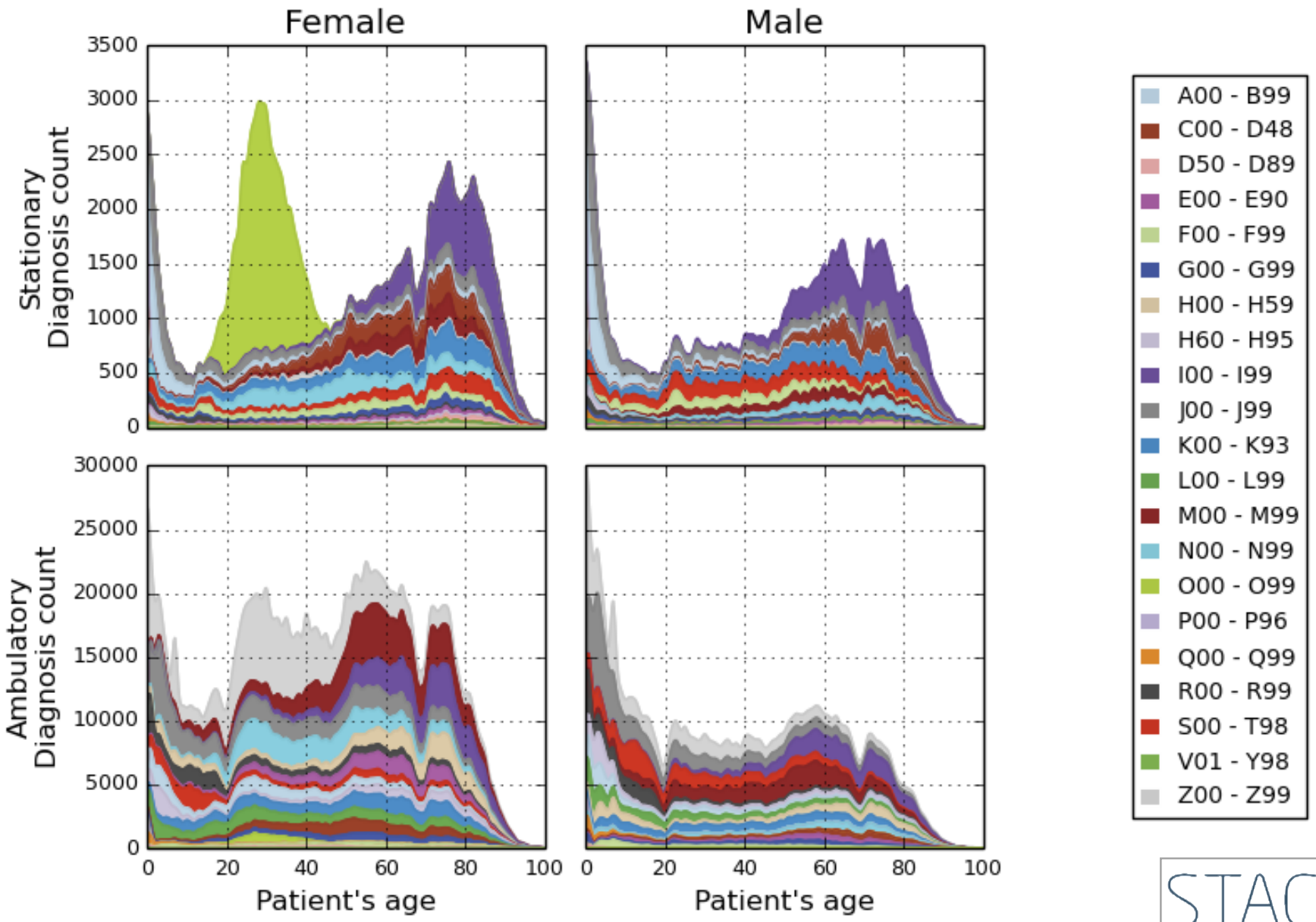

Biokeemilised uuringud (umol/L)
18
16
14
12
10
8
6
17.1
-inf
2006
2007
2008
2009
2010


Biokeemilised uuringud (mmol/L)
12
10
8
6
4
8.8
-inf
2006
2007
2008
2009
2010


Biokeemilised uuringud (mmol/L)
1.5
1.4
1.3
1.2
1.1
1.0
1.29
1.17
2006
2007
2008
2009
2010


Biokeemilised uuringud (g/L)

General overview of diagnoses: Estonia 2012-2013

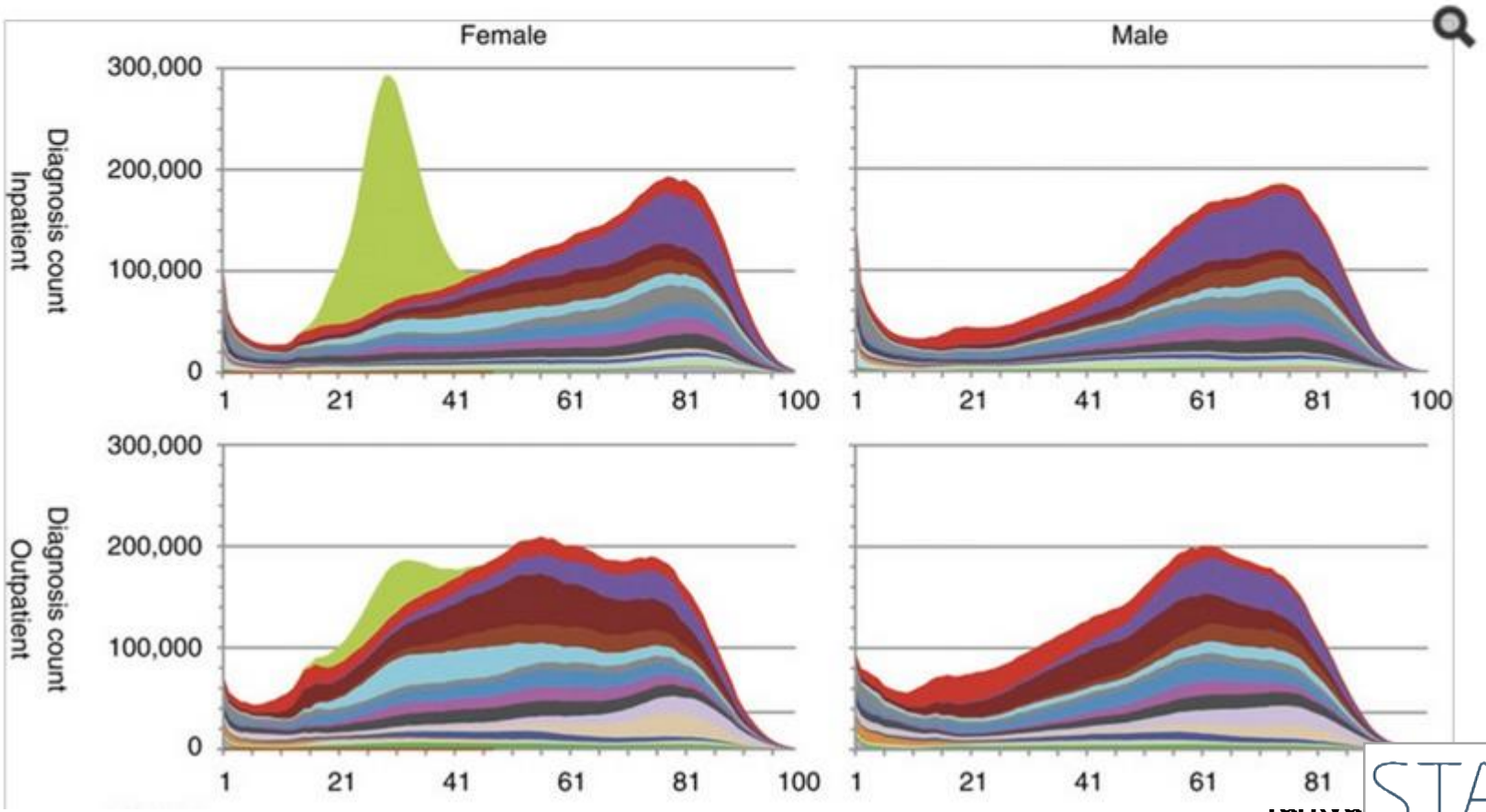


STACC

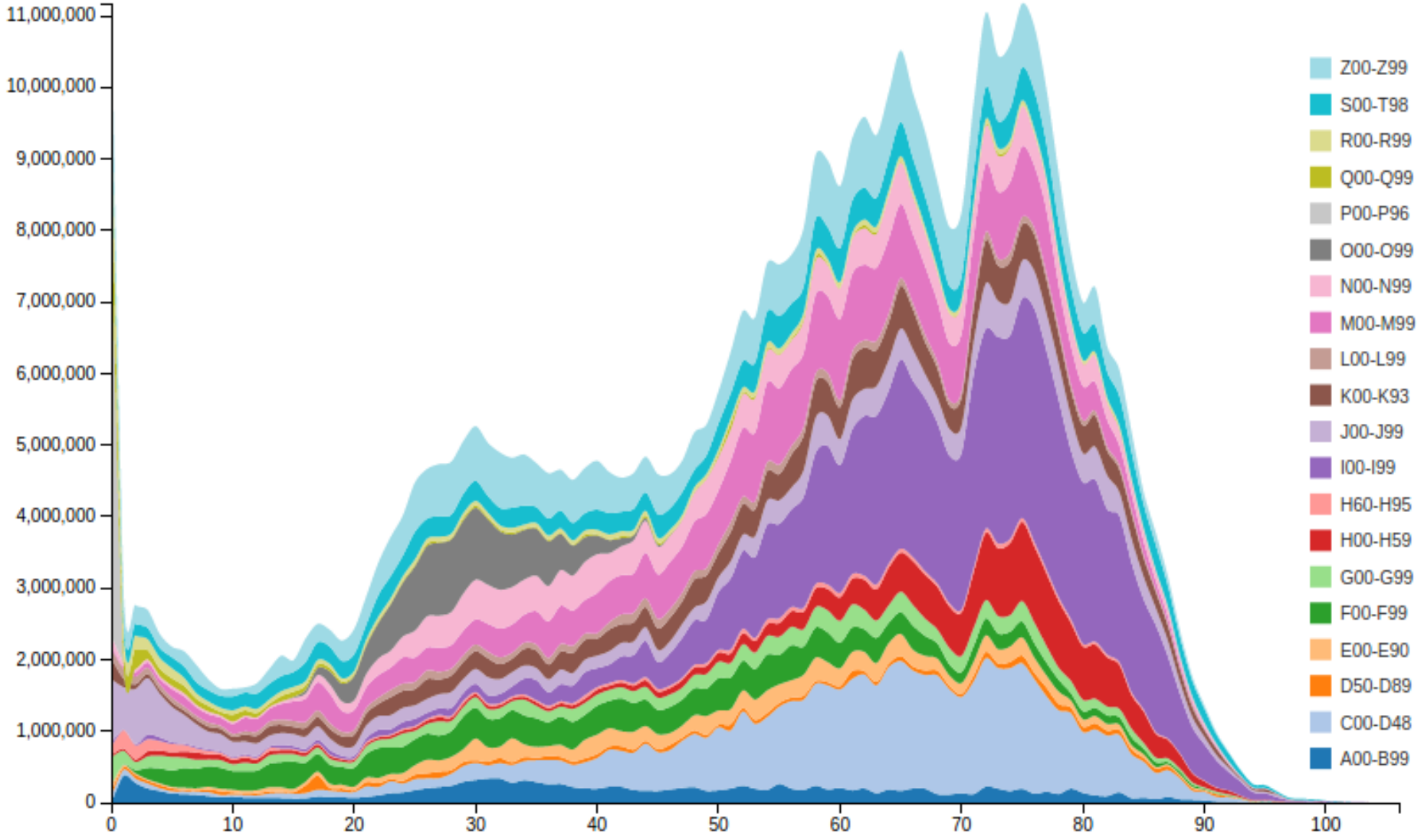
NATURE COMMUNICATIONS
An open access paper
www.nature.com

PMC full text: Nat Commun. 2014 Jun 24; 5: 4022.
Published online 2014 Jun 24. doi: 10.1038/ncomms5022

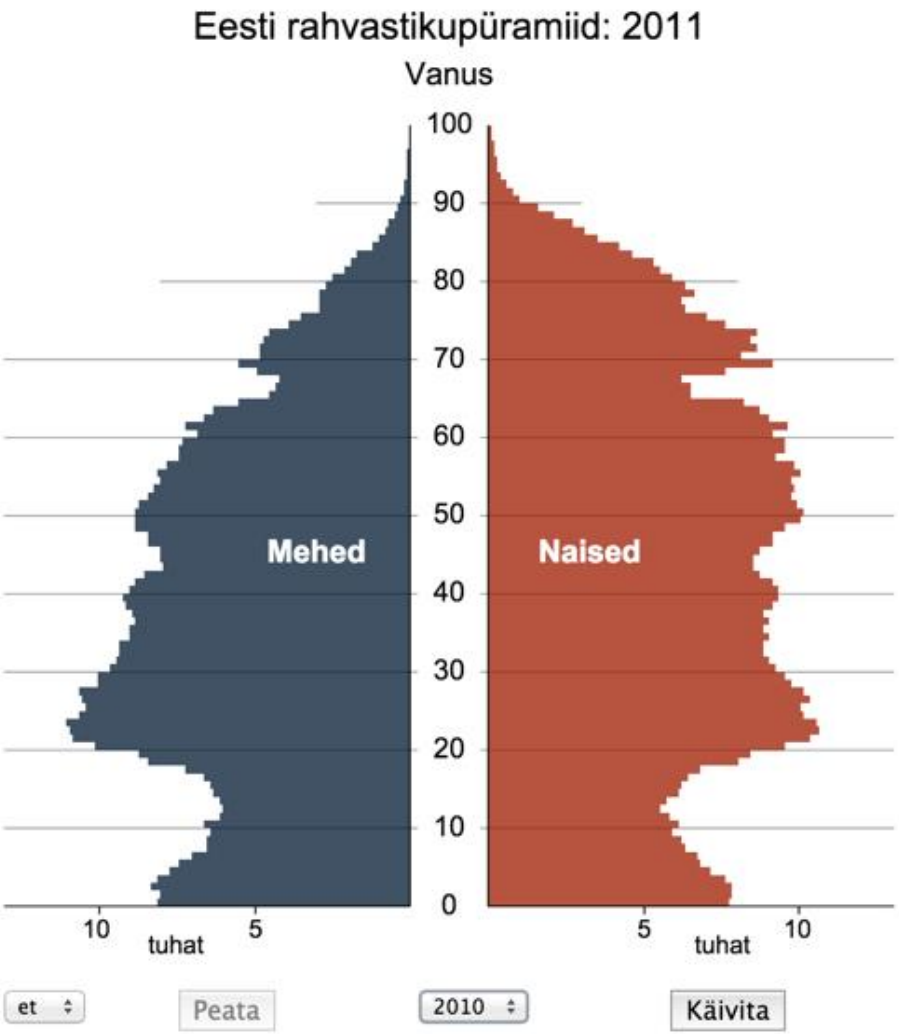
Figure 1



Cost on national health insurance



Tarkvara TAK (STACC)



ES

Rahvaarv

1923 kuni 2011: arvestuslik rahvaarv
2012 kuni 2050: prognoositav rahvaarv

Variant 2 Eeldused:

- summaarne sündimuskordaja tõuseb pidevalt ja jõuab 2047. aastaks kahe lapseni naise kohta;
- suremus väheneb;
- oodatav eluiga sünnimomendil pikeneb 2050. aastaks naistel 80,44 ja meestel 78,36 aastani;
- rännet ei toimu või sisseränne tasakaalustab väljarände.

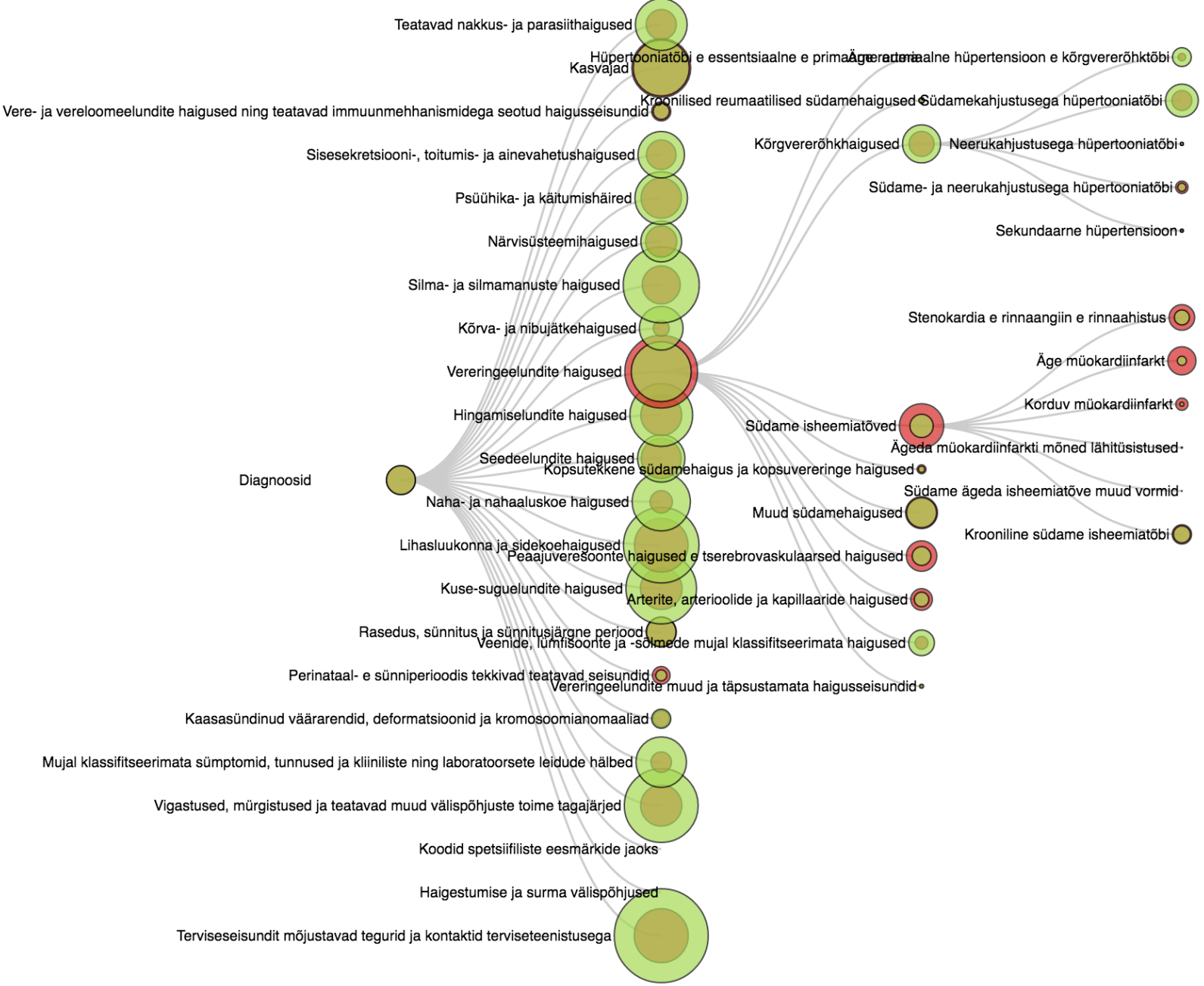
Vanuserühmad

<20	20–64	65+	Kokku		VSM*
277 300	816 300	227 400	1 321 000		28
21	62	17	100	%	

* Vanema vanuserühma inimeste arv 100 keskmises vanuserühmas inimese kohta.

☐ Muuda vanuserühmi
☐ Näita meeste ja naiste arvu erinevust

Tarkvara TAK (STACC)



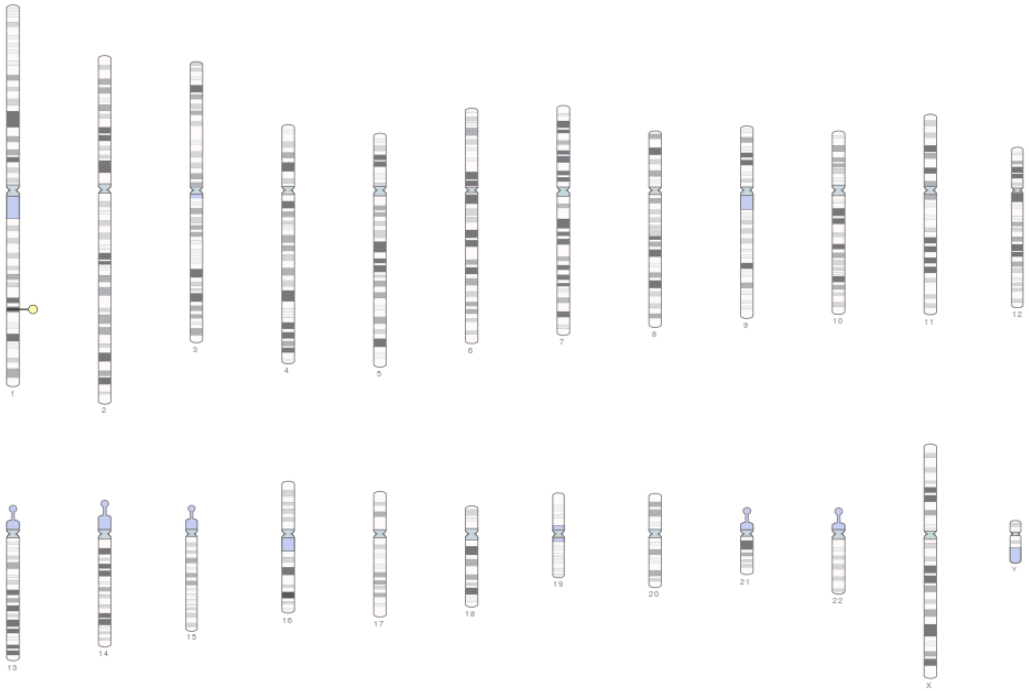
STACC Software Technology and Applications Competence Center

Euroopa Liit Eesti

STACC

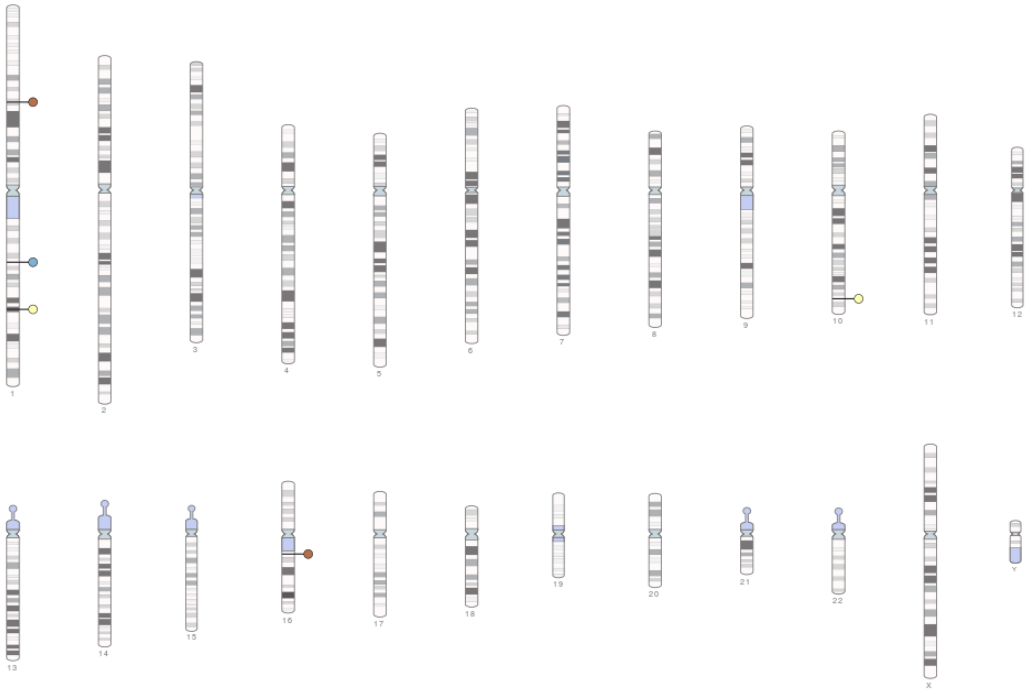
Method	Sample size
Whole genome sequencing	2,400
Whole exome sequencing	2,500
Genome-wide genotyping arrays	20,000
Genome-wide methylation arrays	500
Genome-wide expression arrays	1,100
mRNA sequencing (on-going)	800
Total RNA sequencing	50
Metabolomics (NMR)	11 000
Metabolomics (MS/MS)	1,100
Telomere length	5,200
Clinical biochemistry	2,700

2005



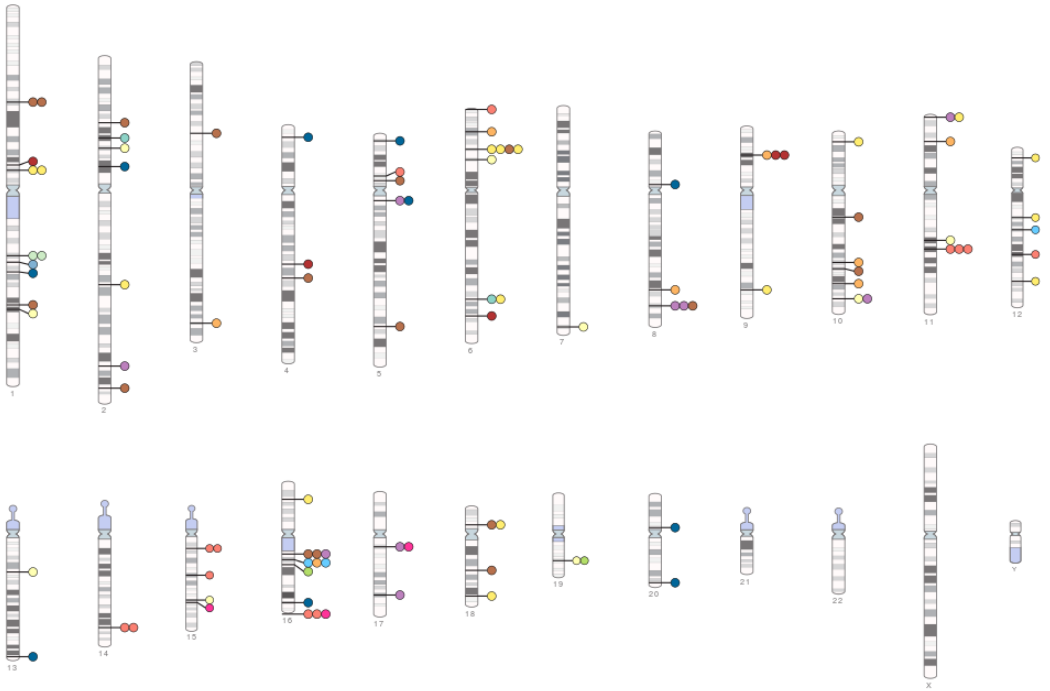
estonian genome center
university of tartu

2006



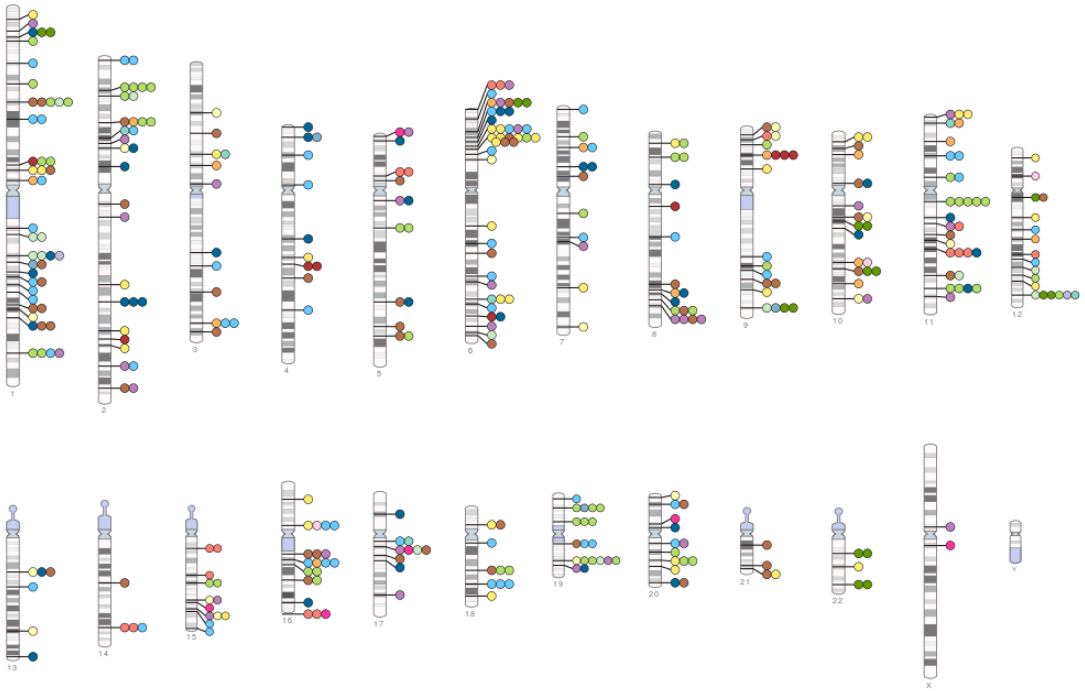
estonian genome center
university of tartu

2007



estonian genome center
university of tartu

2008



estonian genome center
university of tartu

2009

estonian genome center
university of tartu

2010

estonian genome center
university of tartu

2011

estonian genome center
university of tartu

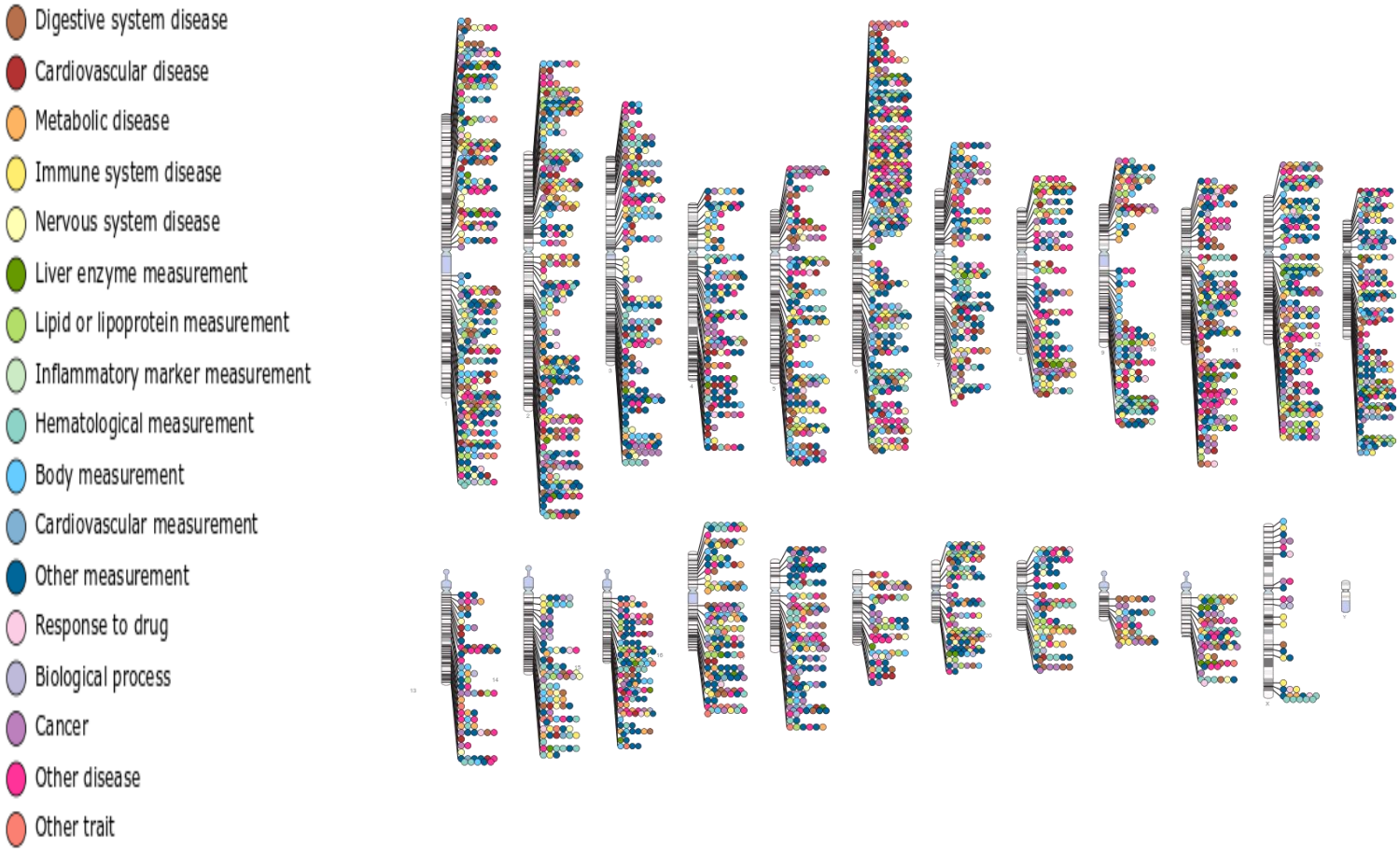
2012

estonian genome center
university of tartu

2013

estonian genome center
university of tartu

Currently


Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait

New Results

Personalized Risk Prediction for Type 2 Diabetes: the Potential of Genetic Risk Scores

Kristi Lall, Reedik Magi, Andrew Morris, Andres Metspalu, Krista Fischer

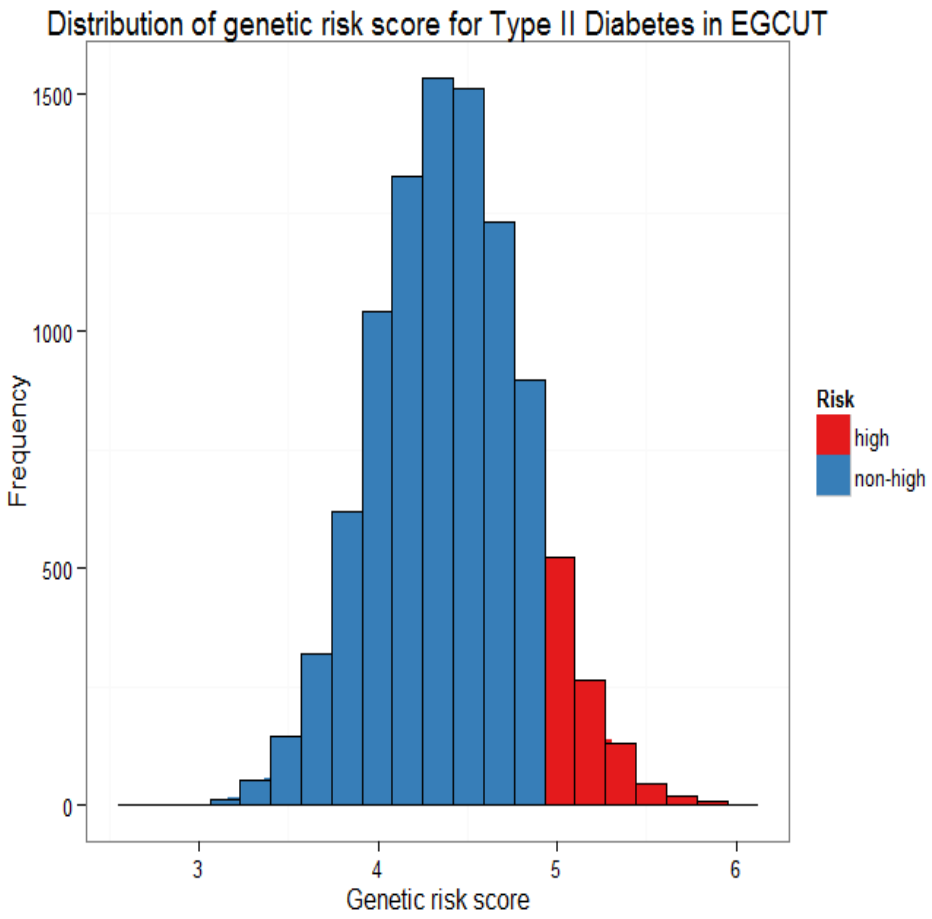
doi: http://dx.doi.org/10.1101/041731

Genetic risk scores

$$\boldsymbol{score = \beta_1 * snp_1 + \beta_2 * snp_2 + \cdots \beta_n * snp_n}$$

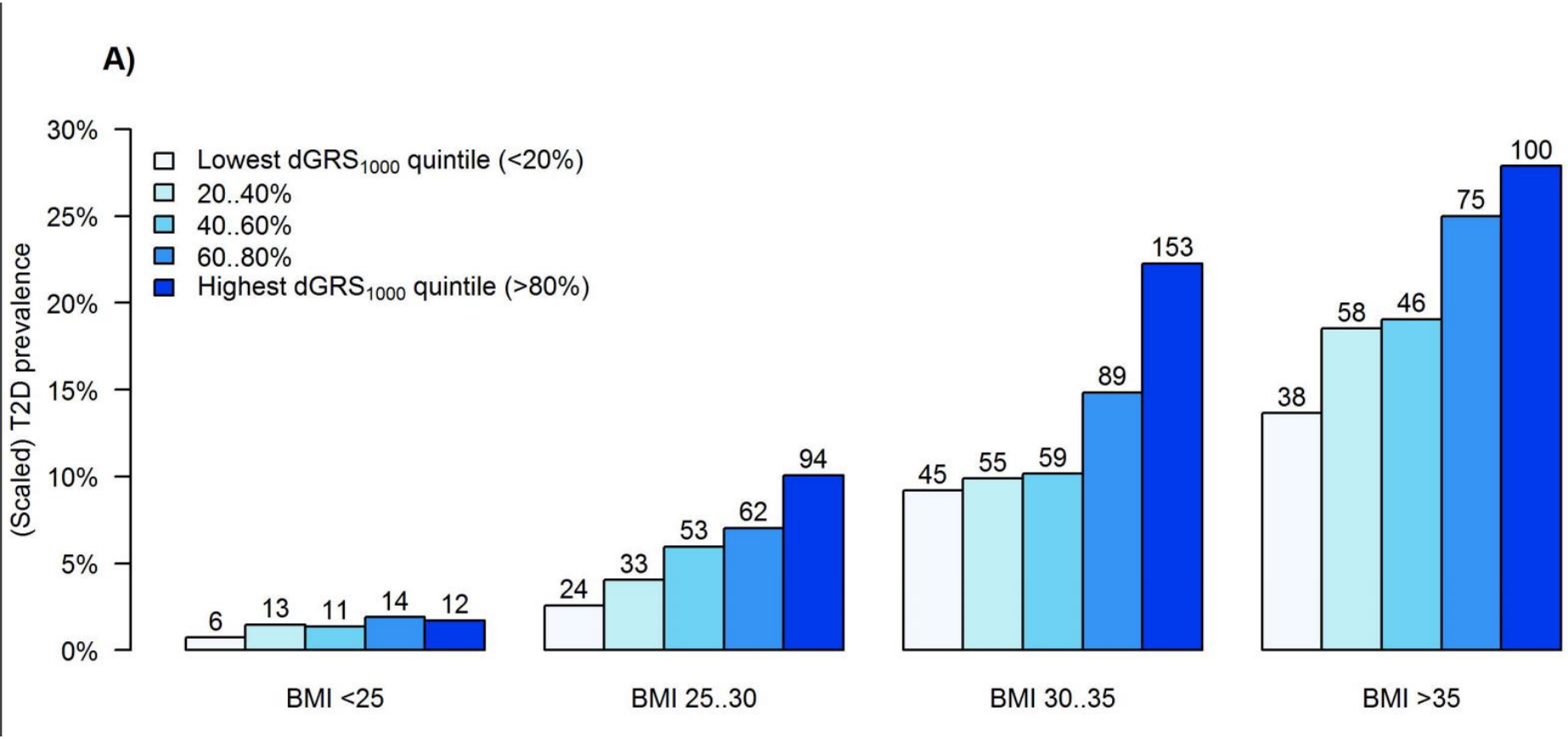
where $\beta_1, \ldots \beta_n$ are known effects from metaanalysis and snp_i is coded as number of risk alleles.

Distribution of T2D polygenic risk score

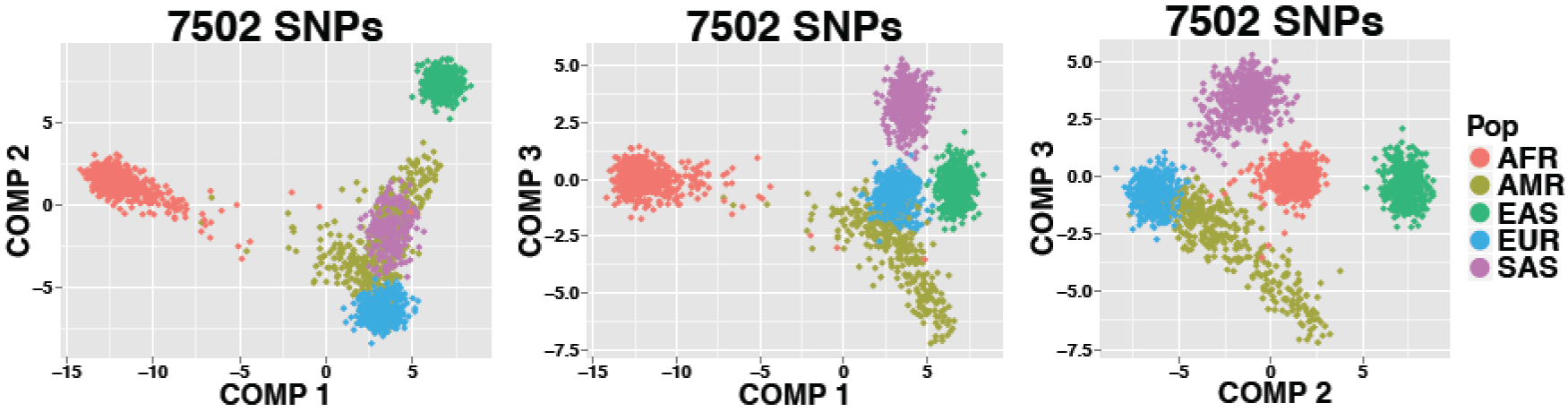


estonian genome center
university of tartu

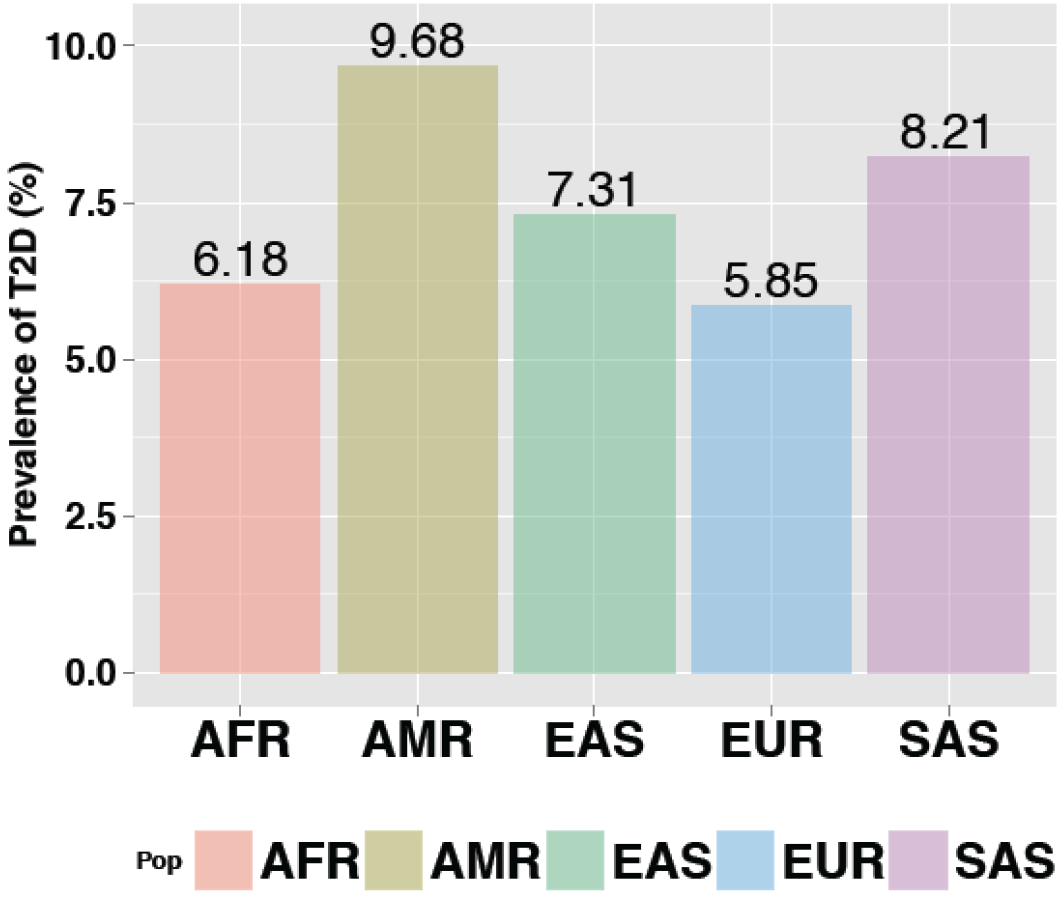
T2D – new incidents based of GRS group and actual BMI



Populations differ based on the same GRS SNPs

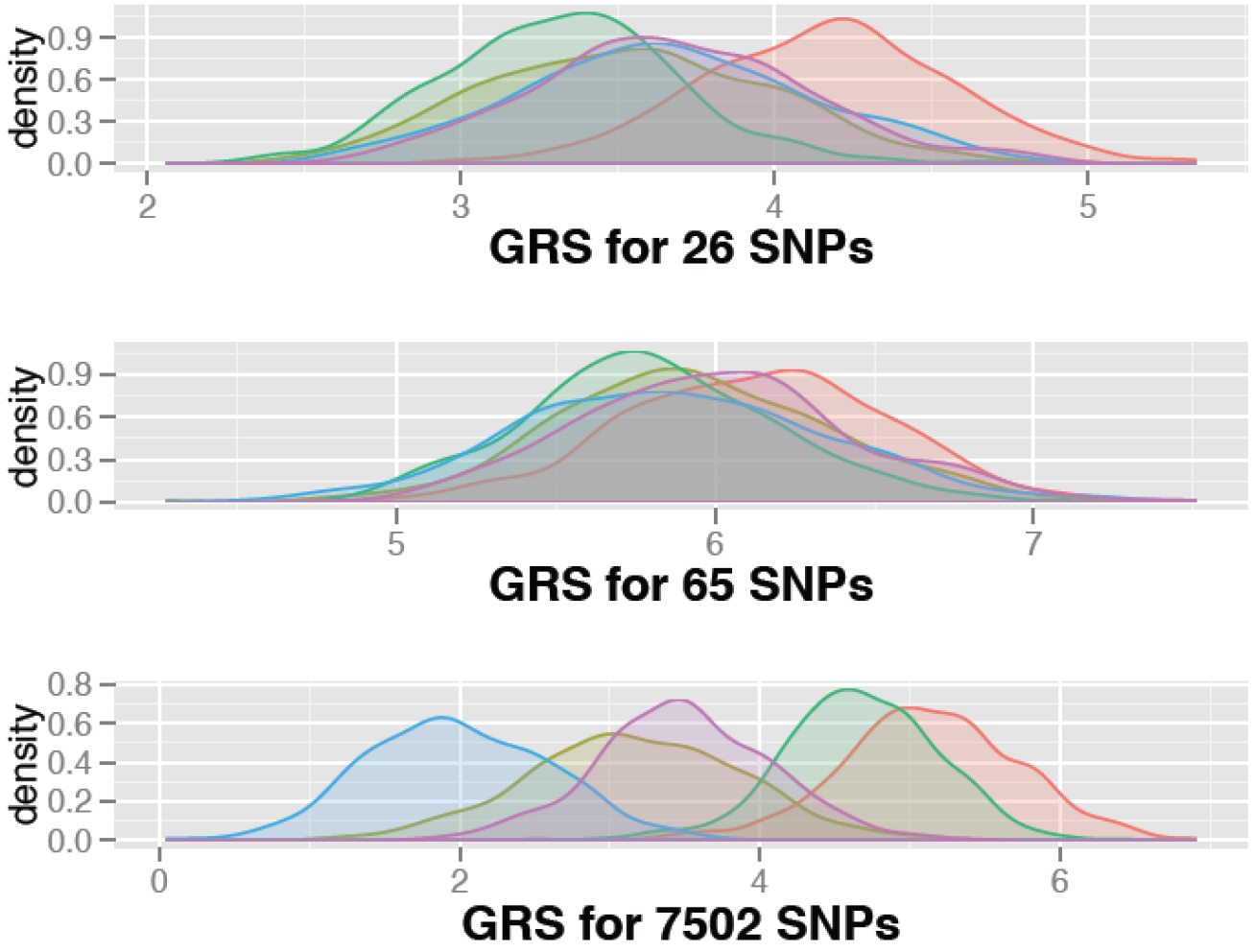


Prevalence[1] of Type 2 Diabetes across five super populations



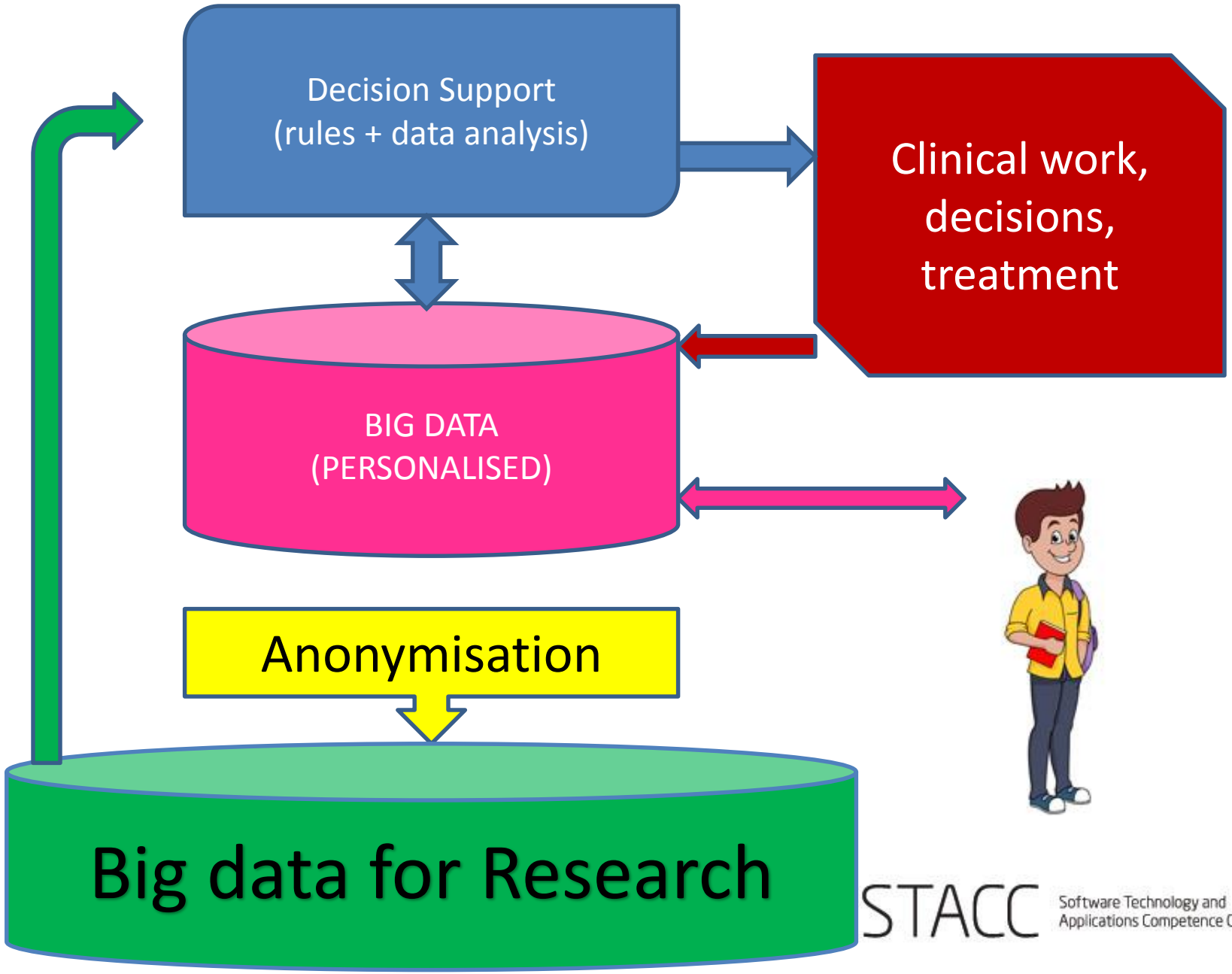
[1] Prevalence is estimated based on data from International Diabetes Federation, 2014, *http://www.idf.org/diabetesatlas*

Comparison of *GRS* distribution between five super populations



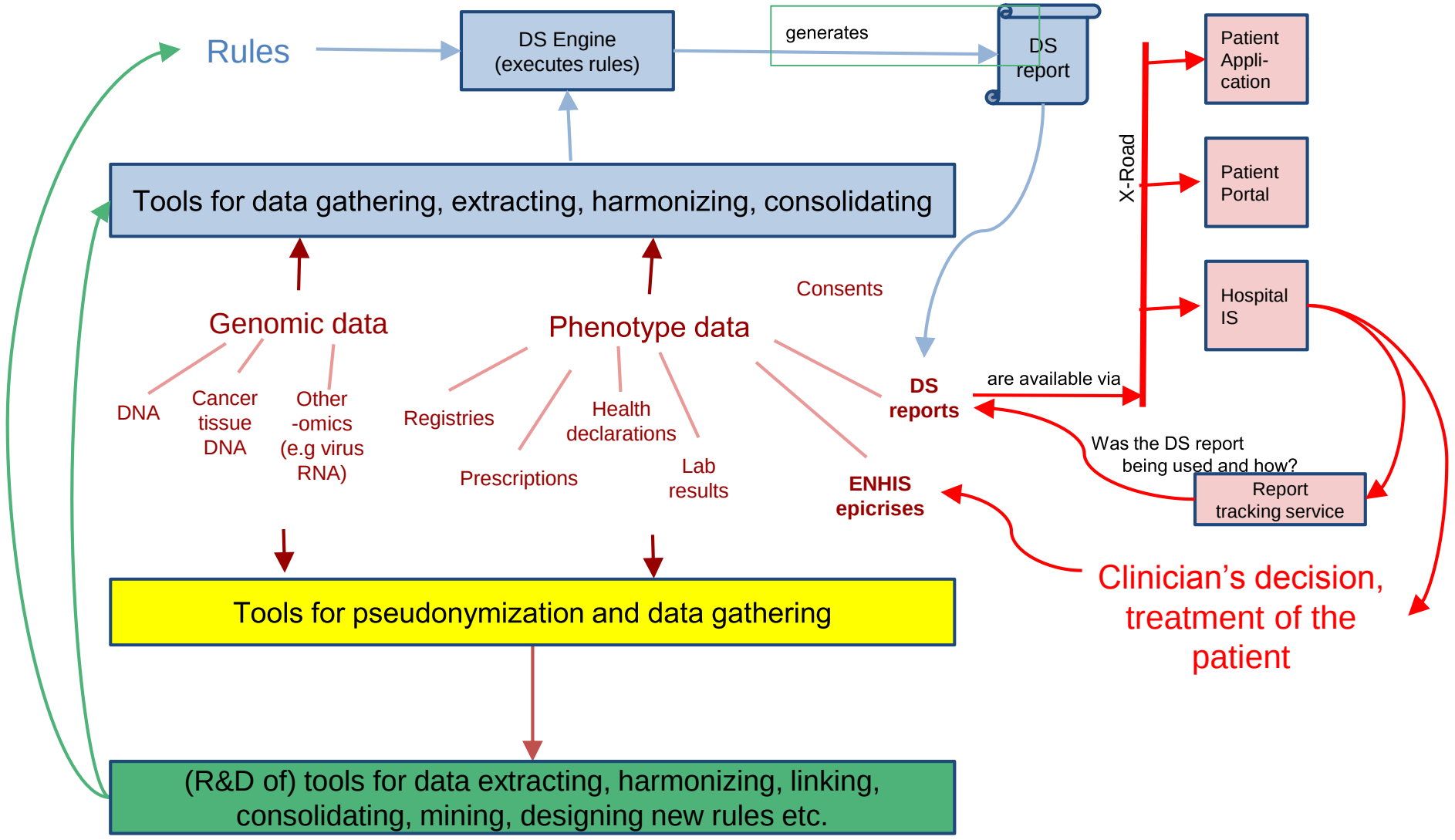
UNIVERSITY OF TARTU

Self-monitoring



STACC Software Technology and Applications Competence Center

What tools & data in each process



STACC Software Technology and Applications Competence Center

Conclusions

- **Electronic health data** analysis is needed for research to reach 4P personalised medicine

- A broad range of data can be obtained nowadays **– secure linking to a person and safeguarding of data is needed**

- **Genetics** – we need good databases of annotated genetic variants and actionable validated predictive models

- **Estonia is an excellent testbed for a Living Lab**
  but lacks a bit of resources

UNIVERSITY OF TARTU

STACC

Software Technology and
Applications Competence Center

estonian genome center

BBMRI-ERIC

Biobanking and
BioMolecular resources
Research Infrastructure

ITEE Centre of Excellence
Digital Connected Economy

it.ee

ITEE Centre of Excellence
Digital Connected Economy

it.ee

UNIVERSITY OF TARTU

TALLINN UNIVERSITY OF TECHNOLOGY

THE UNIVERSITY of EDINBURGH


Thank you!
Jaak Vilo
vilo@ut.ee
TARTU ÜLIKOOL
UNIVERSITAS TARTUENSIS