Personalized medicine

Bioinformatics, fall 2016

Sulev Reisberg

20 October 2016
About myself

Software Analyst
@ Qure TEC

Data analyst and Project Manager
@ STACC
Software Technology and Applications Competence Center
Making big health data analytics (on Estonian central e-health database, insurance fund database, ...) and developing relevant tools for personalized medicine

Data custodian
@ EMIF
Representing Estonian genome center of University of Tartu
Topics

1. **Definition** of personalized medicine
2. Main focus areas
3. Personalized medicine **in Estonia** & what we do at STACC
4. Main **challenges**
5. Homework
What do you think, what is a personalized medicine?
What do you think, what is a personalized medicine?

Precision medicine

Targeted therapy

Stratified medicine

P4 Medicine: Personalized, Predictive, Preventive, Participatory
Definition in Estonia

Personalised medicine refers to prevention, diagnosis and treatment of health disorders, based on individual risk-tailored approach using computational decision support analysis of person’s phenotype and genotype data.

The goal of personalised medicine is to contribute towards preventive, predictive and participatory health system.

Source: Ministry of Social Affairs, 2015
Key drivers (prerequisites) for personalized medicine

- Drop of **DNA sequencing cost**
- Increase of **computational power**
- Rise of **big data analytics**
- Electronic health records (**EHR**)  
- Huge cost and **high failure rate of clinical trials**
- Devices for ‘quantified-self’ (**wearables**)
- **Population ageing**, increase of healthcare costs
- ...

“The **UK** is set to become the world leader in ground-breaking genetic research into **cancer** and **rare diseases**, which will transform how diseases are diagnosed and treated, thanks to a **package of investment worth more than £300 million**, the Prime Minister will announce today.”

David Cameron, Prime Minister of UK

1 Aug 2014

“I want the country that eliminated polio and mapped the human genome to lead a new era of medicine – one that delivers the right treatment at the right time.”

Source: http://www.reuters.com/article/us-usa-obama-genomics-idUSKBN0KU06L20150121
“Estonia has a unique opportunity to be among the leaders of the developers and practitioners of the personalized medicine solutions.”

Main focus areas
Focus area 1:
Pharmacogenomics (PGx)
The main driver for pharmacogenomics

Percentage of the patient population for which a particular drug is ineffective (in 2001)

<table>
<thead>
<tr>
<th>Drug</th>
<th>Efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Depressants</td>
<td>62 %</td>
</tr>
<tr>
<td>Asthma</td>
<td>60 %</td>
</tr>
<tr>
<td>Diabetes</td>
<td>57 %</td>
</tr>
<tr>
<td>Arthritis</td>
<td>50 %</td>
</tr>
<tr>
<td>Alzheimer</td>
<td>30 %</td>
</tr>
<tr>
<td>Cancer</td>
<td>25 %</td>
</tr>
</tbody>
</table>

Drug does not work

Picture source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, ”Clinical Trends in Molecular Medicine, Volume 7, Issue 5, 1 May 2001, Pages 201-204
Cancer means abnormal cell growth regulation.

The main problem: how to make the drug to distinguish such cells from the others.

Targeted chemotherapy: killing cells that divide rapidly and/or block the cell growth.

The selected therapy is usually based on the nature of the particular cancer cell (has somatic mutations, its DNA is different from patient’s healthy DNA etc.).
Pharmagogenomics - choosing right medication (1)

Used to treat breast cancer if the cancer cell is growing because of the estrogen hormone (it has estrogen receptors) (70%)

This drug **blocks hormone receptors** of the cell, preventing hormones from binding to them.

Approved already in 1972
Used to treat **breast cancer** if the cancer cell is overexpressing **HER2 protein** (20%)

More of a protein HER2 (human epidermal growth factor receptor) causes these cells to grow and spread faster than the ones with normal levels of the protein.

**Herceptin** **binds to HER2 receptors** and **blocks them from receiving growth signals**.
Personalized drug selection/dosage

The right drug and right dosage are selected based on the patient’s genome.

Goal: to avoid taking drug that is not working on that patient or causes adverse side effects.
Used to treat HIV

Side-effect:
causes life-threatening hypersensitivity reaction for patients having any *57:01 variants in immune system gene HLA-B
Pharmagogenomics - adjust the dosage

Warfarin: Reduces blood clots and thereby reduces the risk of heart attack

The right dosing depends on several genetic variants (CYP2C9, VKORC1):
break it down too quickly and it will not have the desired effect;
too slowly, and the risk of bleeding increases.
Focus area 2:
Genetic testing related to having a baby
Having a baby

A couple wants to have a baby

Pregnancy

Birth of the child

Genetic testing:

a) Carrier testing

b) Pre-natal diagnostics

c) Early post-natal testing if there are complications
to test, whether you and your partner both have the same recessive abnormal gene (that means: no symptoms) that may result in affected child.

Typically tested diseases:
Cystic Fibrosis, Sickle Cell Disease, Neurodevelopmental disorders and other Mendelian (one-gene) diseases

Usually done in case there has been the disease in the family tree.
b) Pre-natal testing

**Goal:** To detect genetic problems of the baby before birth.

Typically screening for the **most common chromosomal disorders:** Down syndrome (21th chr), Edwards syndrome (18), Patau syndrome (13), also for rhesus factor conflict

c) Post-natal genetic testing if there are complications

**Goal:** To detect, whether the complications are caused by the genetic disorder (and particularly which disorder)

- **Used, if genetic disorder is suspected.**
- **In case** the particular genetic disorder is found (not often), then:
  - carrier screening is performed for the parents usually to detect who is the carrier.
    - **Often (ca $\frac{2}{3}$ cases) none of them is the carrier.** Meaning that the gene mutation was occurred randomly only on the baby (called **de novo mutation**).
Focus area 3: Prevention and early diagnosis
Cancer screening

- There are lots of **gene mutations** that are **associated with high risk of cancer:**
  - Most widely known:
    - some specific mutations in **BRCA1** or **BRCA2** genes lead to **55-65% risk** of developing breast cancer
**Risks for complex diseases**

- Based on **risks scores** that are calculated on **number of different genetic markers**.
  - E.g. risk score for Type 2 Diabetes based on 1000 markers
  - Often combined with other parameters, e.g. body-mass index (BMI)
Focus area 4:
Clinical decision support systems (CDSS)
Current trends of CDSS

Ensuring data interoperability: data standards...

Genetic information added, “quantified-self” data added

Integrating alerts seamlessly into workflow

Shared knowledge-base, accessible via web-services.

From rule-based knowledge-base towards machine-learning based knowledge

Personalized medicine in Estonia
Personalized medicine in Estonia

Estonian Genome Center

- **Biobank,**
  - holding **biosamples of 5% of Estonian population**
    - Including 2300 full DNA sequences and more
    - + **rich health information** (continuously updated) for all participants
- **Researchers**
- Together with Tartu University Hospital carried out **carrier screening, pre- and post-natal testing** (by now these analyses have moved to separate institution Geneetikakeskus)
Personalized medicine in Estonia

Bioinformatics, Algorithmics and Data Mining Group

Bioinformatics research
(incl multi-omics research)

Data mining of Electronic Medical Records,
Wearables (quantified-self), Decision support systems
Personalized medicine in Estonia

+ collaboration with medical doctors
Organizing the medical data

Data analysis

Build analysis tools!

Re-using the analysis outcomes

Building database for effective analysis

Visualizations, decision support systems
Building a database

- Medical records
- Insurance bills
- Data from wearables

Parsing, fact extraction, harmonization, aggregation

Beautiful database
**Sihtasutus Tartu Ülikooli Kliinikum**

**Tegevuslind (litsentsid) nr.:** L01209, L00707, L01325, L00708

**Adress:** Puusepa 1a 50406 Tartu
**E-post:** kliinikum@kliinikum.ee

**Kellele:** Perequisite

**04.05.2009**

---

**Ees- ja perekonnanimi:** MIISU KIIISU 30003012716

**Isikukood:** vanus: 109 aastat

**Haiglas liikumine:**

<table>
<thead>
<tr>
<th>Osakond, allüksus</th>
<th>saabus</th>
<th>lahkus</th>
<th>voodipäevi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hematologia-onkoloogia kliinik, kirurgilise onkoloogia statsionaar</td>
<td>27.10.2008 11:24</td>
<td>27.10.2008 13:44</td>
<td>1</td>
</tr>
</tbody>
</table>

**Voodipäevi kokku:** 2

---

**Lõplik kliiniline diagnoos:**

**Põhidiaagnost:** Quadrans superioexternus mammae; p T4B(2) N1A M0 G1 IIIBst C50.4

**Kaasuv haigus:** Täpsustamata rauavaegusaneemia D50.9

---

**Annamees, diagnoosi põhjendus ja haiguse kulge:**


---

**VASTUSED:**

Hemogramm kolmeosalise leukogrammiga

- WBC 6.2 (4 .. 10 E9/L)
- RBC 4.37 (3.8 .. 4.8 E12/L)
- HGB 123 (120 .. 160 g/L)
- HCT 39 (36 .. 47 %)
- MCV 88.8 (83 .. 101 fl)
Typical EHR document in Estonia (HL7v3)

14 such documents for every Estonian (on average)
HL7v3 XML example 1

...
HL7v3 XML example 2

...<title>Anamnees, diagnoosi põhjendus ja haiguse kulg</title>
...<text><list ID="ANALYSIS"><item ID="ANA"><caption>Anamnees</caption><content>
Et, hospitaliseeritud dr. Krafti poolt täiendavates uuringutes. Tegu viie sageli süüdane vahelööke, vahel hakkab rooldsas seisminassse olnud.  
Paar käitumistes on tõenäoline stenokardiaat ei ole olnud ja ka
peale seda hübriidne arvamus kehtestati.  

Vererõhkk varem alati normis olnud, viimase ajal veidi kõrgem.  

...</content></item></list></text>

Complaints, observations in free text parts

...<entry typeCode="COMP">
<procedure moodCode="EVN" classCode="PROC">
<code codeSystemName="Etagi" codeSystem="1.3.6.1.4.1.1028.1.2.1.6.1" code="6326" displayName="Holteri
monitoring" />
</procedure>
...<text>Billiske, Svea:  
Jälgitud periodil esines sündmus. 7714
polümorfast VES (1%), naist 201 VES paari, 2 VES grupp (3 QRS kompleksi järjest), 60 isoleeritud EVES. 1 SVES paari, 5 SVES grupp (3-8 QRS kompleksi järjest). Aistooloosid pause ei esinevad.  
Max. fr. 122 x/min k.15.53., min. fr. 41 x/min k.15.47., keskmine fr. 62  
</text>

Pulse measurements in free text parts

...<procedure moodCode="EVN" classCode="PROC">
<code codeSystemName="Etagi" codeSystem="1.3.6.1.4.1.1028.1.2.1.6.1" code="6326" displayName="Veloeugmeetría" />
</procedure>
...<text>Abramova, Lidia:  
Saavutades Max. Laad: 125 Watt. Rahuleku pulsiaegude 75 x min, koormusel maksimaalne pulsiaegude 142 x min. Patsient saavutas 86 % elülisest maksimaalset pulsiaegudest. Rahuleku RR 100/70 mmHg, maksimaalne RR 140/80 mmHg  
</text>

Blood pressure measurements in free text parts

Tarkvara TAK (STACC)
Height, weight and BMI in free text, with error (swapped values)
Anonymization tool

**Input text:**

Patsient John Doe Vanus 44 a. IK – 77771478888 võeti statsionaarsele ravile.

**Anonymized text:**

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

<table>
<thead>
<tr>
<th>Original data</th>
<th>Extracted data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siinusrütm 59 lööki min</td>
<td>59</td>
</tr>
<tr>
<td>Kiirenenud, Ekg-l siinus rütm, 160/min</td>
<td>160</td>
</tr>
<tr>
<td>vatsakeste tahhüarütmia fr - ga 150 lööki min</td>
<td>150</td>
</tr>
<tr>
<td>EKG --&gt; siinusrütm, fr. 110 x min</td>
<td>110</td>
</tr>
<tr>
<td>Ps 66/min</td>
<td>66</td>
</tr>
<tr>
<td>Fr = 77 ' min</td>
<td>77</td>
</tr>
<tr>
<td>Maksimaalne fr=196, minimaalne fr=55</td>
<td>196; 55</td>
</tr>
<tr>
<td>V/v-l RR 133/105/90 , arütmia</td>
<td>90</td>
</tr>
</tbody>
</table>
Detecting the meaning of abbreviation

Context

... p silm ei näe ...
... kolmas p palavik ...
... vähene p pleurareaktsiooni riba ...
... p 6mm ümmargune ...

<table>
<thead>
<tr>
<th>Full form</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parietaalne</td>
<td>92%</td>
</tr>
<tr>
<td>Parem</td>
<td>4%</td>
</tr>
<tr>
<td>Päev</td>
<td>3%</td>
</tr>
<tr>
<td>Pupill</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
**TEXTA** - for finding contextually similar documents

**Enter seed words:**

- paracetamool
- ibuprofen

**System provides contextually similar words:**

- paracetamool
- ibuprofen
- ibuprofeni
- paracetamol
- paracetamool
- ibumetin
- paracetamo
- paratsetamol
- ...

This helps you to find **more documents**:

- tegemist on tõenäolisel t kõrgerverõhutõvega, soovitat raviskeemi korrigeerida. Vajadusel paracetamol, ibuprofen. Vaevust püsimestel - kordu konsultatsioon, siis toob uuringu vastust.
- PER parema sääre ja reie välispinnal, eelnevalt olnud mingi muhk mall lat kohal. Ravimeits kasutanud paracetamoli ja lokaalselt ibuprofeni.
- 06.03.2013 - Suunatud perearsti poolt. Õhurõhutised, ülakõhuvärvused, omeprazolum olulise efektita. Peavalude pärast kasutab ibuprofeni. Appendiitoomaa, 3 Keissri lõikus.
- 11.04.2013: ei ole sugugi parem. Sovit paracetamoli 2tx4, ibuprofeni 400x4, HI pikendus, reziiim
- 17.04.2013: on parem, ravi jätakub
- 19.05.2011 12:44

  19.05 Kaebused tugevad valud seljas nimmeristljuu piirkonnas ja irred parem jaaljas juba 2 kuu.O-b Fisioloog lordoosis lamen Kummardamine ette poole valuik ja piiratud Reff ahhil abs deextra Hypeestis S1 sinistra Määrata KT ja Tab ibuprofeni 0.4 Tulla peale KT
- 03.06.2011 iSeleasundi paremnine,antud soovitused,RKK,Paracetamoli gr valudel puhul
- 24.09.2012 15:01 -

  D04743 - E230 - neuroloogia

Getting an overview of the whole database

Budget of Estonian Health Insurance Fund

Cardiology
Budget: 11M
Spent: 13M (115%)
Where the people from Lääne-Virumaa are traveling to in order to get their cataract surgery
Average annual cost of healthcare services of each Estonian citizen by age
Overview of the patient

Diagnosis

Blood pressure

Blood glucose

Cholesterol

Glucated hemoglobin

Diabetes
Overview of the patient

- **Diagnosis**
  - Diabetes

- **Tests and Treatments**
  - Blood pressure
  - Glucose levels
  - Hemoglobin
  - Cholesterol

- **Medical Data**
  - Medical records
  - Wearables
  - Genetic data
  - Medical insurance bills
Decision support system for family doctors
Quantified-self (wearables)
Raw data of pupil footsteps in Estonian primary schools
Average number of footsteps made in different lessons in Estonian primary schools
Main challenges
Challenge 1:
Does personalized medicine have any impact?
28% of health IT solutions have proven some positive impact to the patients.

72% of the solutions provide non-significant or mixed impact for the patient.

One of the first of few studies that show that using personalized (biomarker-based) strategy lead to improved cancer patient outcomes (2015).

Targeted cancer treatment is the most promising field, but...

• Most of the “targeted” drugs only block one of the tumor growing pathways

• Cancer cells almost always develop a resistance to single-targeted drug

• Most of the drugs are too toxic to use in combination
Challenge 2:
How to organize the massive (and growing) amount of data and knowledge?

Genomic basis of the disease
Influences of disorders of translation and expression
Behavior
Environment
Take-home message
Take-home message

Personalized medicine
is more precise clinical approach
by using patient’s genomics
(and also other data)

Main challenge
Show that personalized approach
has a positive impact

Ready-to-use focus areas
Targeted cancer treatment
Pre- and post-natal diagnostics

In Estonia
Yes! We do!
Homework
(2 tasks)
Task 1: analyzing schoolchildren activity

Given: footstep data of 4 primary school pupils: A, B, C, D (measured in Sep 2016)

One schoolday, from 8 till 15 o’clock
NB! Pupils are from different schools!

Your task: analyse the data of 3 pupils according to your matrice number
Task 1: download data

https://www.dropbox.com/s/ma6zwg2v3mlzqkw/pupildata.zip?dl=0

Contains four files: pupila.csv, pupilb.csv, pupilc.csv, pupild.csv

Each looks like this:
Task 1: select pupils

Select pupils from this table according to the last number of your student ID:

<table>
<thead>
<tr>
<th>Last number of your student ID</th>
<th>Pupils to analyze</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1 or 2</td>
<td>A, B, C</td>
</tr>
<tr>
<td>3 or 4</td>
<td>A, B, D</td>
</tr>
<tr>
<td>5 or 6</td>
<td>A, C, D</td>
</tr>
<tr>
<td>7, 8 or 9</td>
<td>B, C, D</td>
</tr>
</tbody>
</table>

E.g B55569
Task 1: Data preparation (1)

a) Take the data of 3 pupils and save it to the same dataset.

b) Get some sort of heatmap of your data. You may use whatever tool you like for this. R is an option. Also Excel with its “conditional formatting” feature works well.

NB! You do not need to report the heatmap in your homework.
c) All lessons last 45 minutes. Two pupils are from the same school and same class and the third one is from another school. Having that knowledge, detect at what times each lesson begins and ends in both schools.
Task 1: Answer to questions (1-5)

1. Which pupil has been the **most active**? Who is the least?

2. Which pupils are **from the same class**? Justify your answer.

3. In each school, there is a **breakfast break** where the food is provided. Can you find, which break is that in each school? How did you find it?
Task 1: Answer to questions (1-5)

4. One or two pupils in your data had the **physical exercise lesson**. Which one(s)? Justify your answer.

5. Imagine that you have 100x more pupils measured this way from different schools. What do you think, what are the **difficulties** of analysing such data?

   Warning: this is a real-life data, it’s quite messy. There are no 100% correct answers in this task.
Task 2: analyzing healthcare bills

Given: 0.1% of all healthcare service bills in Estonia in 2015

Your task: analyse the data
Task 2: download data

https://www.dropbox.com/s/jorfgtk2yqqfqto/healthcare_services_2015_sampling.csv?dl=0
Contains 0.1% of healthcare service bills in Estonia 2015.

Looks like this:

```
icd10  cost
A04  1.4
A04  2.5
A04  2.4
A08  5.6
A08  5.4
A08  2.6
A08  2.5
```

Each row is one bill.

Each bill contains:
* Diagnosis code (ICD-10 classification)
* cost (euros)
Task 1: Answer to questions (1-2)

1. Find 3 bills with the the largest costs. What are the corresponding diagnosis names? Use http://apps.who.int/classifications/icd10/browse/2016/en to find the names.

2. When summing the bills with the same codes together, what is the most expensive diagnosis code? What is the corresponding diagnosis name? How large amount of total costs it uses?
Send homeworks to sulevreisberg@gmail.com by 27th October 12:00 (1 week!) (PDF please)! Thank you!