Sensitive Data Challenges & Beacon as a Discovery Tool

Babita Singh
Postdoc Bioinformatics
EGA-CRG, Barcelona - Spain
babita.singh@crg.eu
Human Genomics Data
Data: the good the bad and the ugly

AI systems are intrinsically tied to human bias — biased and incomplete data

Medical data for training is (often) characterised how?
- mostly white
- mostly male

Straw I, Wu H, Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction, BMJ Health & Care Informatics 2022

44% cases of liver disease among women missed!

What is our knowledge on viral sequences?
- very incomplete
- skewed to human-infecting viruses
- strongly imbalanced relative to the host sequences


Traditional (ML and non-ML) methods poorly generalise to OdD

Explicit or hidden class imbalance implies reasoned choices of data, models and metrics

Thökö et al., Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data, bioRxiv 2022

12 September 2022 / Workshop on Machine Learning Good Practices
Challenges with HealthCare data

Human Genomics - changing environment

Percentage human genomes and exomes that are funded solely by healthcare systems

- 2012: 1%
- 2017: 20%
- 2022: 80%

CHALLENGES

- Data still geographically distributed
- Dynamics of how we access data will change
- Clinical data are not interoperable
- Healthcare is not used to this type and amount of data: terabyte to exabyte
- Technical knowhow is in the research community
- Attitudes and action towards open data need to progress
- Secure access and governance

Challenges with HealthCare data

1. Geographical distribution of data
2. Data access dynamics will change
3. Clinical data is not interoperable
4. Not equipped for high volume of data
5. Technical challenges from data generation to research
6. Attitude and action towards open data
7. Secure access and governance

Genomics and HealthCare data and European Genome-phenome Archive (EGA)

1. Geographical distribution of data
   FEGA/Local EGA

2. Data access dynamics will change
   DUO/AAI/Data passport

3. Clinical data is not interoperable
   GA4GH tools & standards

4. Not use to high amount of data
   EBI storage, Partnership with BSC, technical expertise

5. Technical challenges from data generation to research
   Data discovery Tools development

6. Attitude and action towards open data
   Active Collaborations, Reach-out activities

7. Secure access and governance
   Tools & standards

.. In collaboration with ELIXIR and The Global Alliance for Genomics and Health (GA4GH)
Healthcare Data Discovery
Healthcare Data Discovery Challenges

STXBP1 Genetic Mutation
@STXBP1.Family · Community

Shank2 Gene Mutation
March 23, 2021 at 10:51 AM · 0

Calling all families: The SHANK2 Foundation is in the process of creating our website. We are in need of family stories.

SAMD9L Gene Mutations
Private group

CDK13 - Genetic Mutation
@CDK13Geneticmutation · Community

Cyclin Dependent Kinase 13 Genetic Mutation
Health data in clinics challenges

LipidTech Genetic Testing Report

Patient details

Patient: JAMIE-13
Address: 1 Double Helix Road, Arkadelphia, AR 71923
Sample received: 02 Jan 2020
Report date: 19 Feb 2020

Result of the genetic analysis

Heterozygous pathogenic variant in the LDLR gene.

Result summary

Gene: LDLR (RefSeq NM_000257.4)
Genetic Identifier: c.1946G>A (exon 11)
Mutation class: Amino acid change

Interpretation

This mutation is directly associated with familial hypercholesterolemia, since its pathogenicity has been validated. Validation study: Hibske et al. (1990) Ann Rev Genet 24:133

Recommendations

Referral to a genetic counselor is recommended.

Methods

LipidTech FH genetic analysis detects substitutions and indels in exons and intron-exon boundaries of the following genes:

- LDLR gene (familial hypercholesterolemia): 18 exons
- PCSK9 gene (inhomogeneity): 16 exons
- LTA/AT2P1A (ARID): 37 exons
- APOB gene (familial defective APOB): regions of exons 28 and 29 involved in LDL binding
- APOE gene: region of exon 4

The assay also detects copy number variations (CNVs) in the LDLR gene associated with FH.

Proprietary algorithms are used to detect genomic deletions and duplications (del-dups) in all 18 exons and the promoter region in the LDLR gene.

All genes are fully analyzed by next generation sequencing on the Illumina MiSeq sequencer. 70 regions of interest are amplified from genomic DNA, isolated from whole blood or saliva, in multiplex PCR reactions that include highly purified primers for the specific amplification of regions in which mutations causing FH can be found.

GENOME REPORT

RESULT SUMMARY

The discovery of this alteration in its gene was performed and covered 67% of all patients at 8K coverage or higher, resulting in over 0.5X overreads (reported over the reported genome). These data were analyzed to identify previously reported variants of potential clinical relevance in a subset of panels. The following list of variants was identified for clinical evaluation (methodology): 30% of variants when identified in a subset of patients.

1. RESULT 1 RELEVANT TO INDICATION/TESTING

For patients with a familial hypercholesterolemia, sequencing of this gene will provide valuable information for identifying genetic variants associated with familial hypercholesterolemia.

SECONDARY RESULTS

2. ADDITIONAL MEDICAL INFORMATION

In patients with familial hypercholesterolemia, this variant is not associated with a higher risk of cardiovascular disease.

3. GENETIC RISK ASSESSMENTS

In patients with familial hypercholesterolemia, this variant is not associated with an increased risk of cardiovascular disease.

4. RED BLOOD CELLS AND PLATELET INFORMATION

In patients with familial hypercholesterolemia, this variant is not associated with an increased risk of red blood cell or platelet dysfunction.
Health data in clinics challenges
Health data in clinics challenges

Pilot Project: To connect Hospitals of Catalunya through Beacon

No Guidelines to store and share genomics data
“For example, we have had a lot of problems regarding the data-protection and privacy of data in terms of sharing..

..Even though it is clear that the privacy of the patient is not at risk, it's difficult to understand the legal responsibilities of the different hospitals.

The other thing is the technical limitations of the hospitals. Not everyone has all resources available, and for example, in general, there are no previous databases to start with.”

- Dèlia Yubero Siles, Clinician, Sant Joan de Déu Barcelona Hospital
Possible solutions?
Shifting focus..

Centralised knowledge Discovery

..In this scenario, clinical data does not get shared at all

Decentralised knowledge Discovery

Distributed repositories
“Those who explore an unknown world are travelers without a map; the map is the result of the exploration.”

—HIDEKI YUKAWA, PHYSICIST
THE BEACON PROJECT  version 01 (2018)

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.

.. in collaboration with ELIXIR
Beacon v1 vs. v2

Before

Do you see mutation X at position Y on Chromosome Z?

Yes

Umm.. And??

That’s it.

Say more? Please..

Beacon v1.0

After

Umm.. And??

.. And this mutation was also found in Disease A, B and C

Oh and do you have information about: Phenotype, Studies, Variant annotation?

Yes

Cool, how can I get the data?

Please apply to DAC P to access data for dataset P,Q,R

Wow, thankyou Beacon v2.0 you’re so cool!

Beacon v2.0

 QUERY BY TYPE
 DIFFERENT FILTERS
 NEW SCHEMA VERSIONS
 DIFFERENT ACCESS LEVELS
 BEACON NETWORK INTEGRATION
A mutation in the gene: **dystrophin**, a protein that is essential to the proper functioning of our muscles.
**Beacon Models**

**INDIVIDUAL**
- Gender: Male
- Age: 25
- Ethnicity: ..

**BIOSAMPLE**
- Muscle Fibre
- Origin: Leg

**RUN**
- Platform: Illumina
- Process: Whole genome sequencing

**ANALYSIS**
- Software Pipeline: YYY
- Variant-caller: NNNN

DNA sequence data from an automated sequencing machine

![Sequencing Results](image-url)
Beacon Models

5. COHORT
   Ages: 25-55
   Disease: Dystrophy
   Control: Healthy

6. DATASET
   Disease: Muscular Dystrophy
   Study id: Muscular Dystrophy

7. GENOMIC VARIANTS
   Chromosome: X
   Start position: 2000
   End position: 2001
   Mutation: C > T
   Diagnosis: Muscular dystrophy
   Risk: High

[Diagram of genomic variants and healthy vs. patient DNA strands]
Beaconize your data

* Beacon v2 Reference Implementation *

XLSX | Metadata (incl. Phenoclinic data)

Validation

VCF | BFF | Database | API

Request | Response

MongoDB
Beacon Use-cases

Implement Beacons in Rare-diseases dataset (Funded by LA MARATÓ)

SIX Hospitals of Catalunya

Hospital Sant Joan Déu

Hospital Vall d’Hebron

Hospital del Mar

Hospitals of Catalunya on Beacon Network

Grant: La Marato, 2020
Beacon Network: Aggregator of ALL Beacons
## The Beacon Continuum

### Beacon Type

- **Boolean Beacon**
  - Boolean
  - Counts
  - Example preview
- **Summarized data Beacon**
  - Boolean
  - Counts
  - Example preview
- **Consortia Beacon**
  - Boolean
  - Counts
  - Data preview
  - Full access
- **Internal Beacon**
  - Boolean
  - Counts
  - Data preview
  - Full access

### Response type

- **Handovers and links**
  - Boolean: ✗
  - Counts: ✓
- **Authorised Access**
  - Boolean: ✗
  - Counts: ✗
- **Preview**
  - Boolean: ✗
  - Counts: ✗
  - Data preview: ✓
- **Full access**
  - Boolean: ✗
  - Counts: ✗
  - Data preview: ✓
  - Full access: ✓

### Response type example

- **Boolean Beacon**
  - Yes/NO
  - Example preview
- **Summarized data Beacon**
  - Yes, 123 cases
  - Example preview
- **Consortia Beacon**
  - Example preview
- **Internal Beacon**
  - Example preview
Beacon Query

Show all variants in rare disease X

Public version

⚠️ You have to be logged in to see this result.

Controlled access
### Beacon Query

**Variant Information**

- **variantInternallId**: chr22_16050678_C_T
- **genomicHGVSId**: 22:g.16050678C>T
- **variantType**: SNP
- **position**: 16050677
- **referenceBases**: C
- **alternateBases**: T

**molecularAttributes**

- molecularAttributes: .
- genelids: .
- molecularEffects: intergenic_region
- annotationImpact: .

**CaseLevelData**

- HG03455 0/1
- NA20538 0/1

**Result Type**

- Tidy

**Showing** 1 to 1 of 1 entries
Tooling: Beacon Verifier
Approved by GA4GH: THE BEACON PROJECT version 02 (April, 2022)

..first specification for Genomic data discovery

Jordi Rambla (EGA), 22 April, 2022
Beacon search

TP53 mutations in ovarian cancer stage 4, women >45 age
Genomics Beacons

Internet of Genomics

🔍 TP53 mutations in ovarian cancer stage 4, women >45 age

Beacon search  I'm Feeling Lucky
The Beacon Team

Tony Brookes
Tim Beck
Colin Veal
Tom Shorter

Sergi Beltran
Carles Hernandez

Serena Scollen
Gary Saunders
Giselle Kerry
David Lloyd

Joaquin Dopazo
Javier Pérez
J.L. Fernández
Gema Roldan

Michael Baudis
Rahel Paloots
Hangjia Zhao
Bo Gao

Inserm
David Salgado

H3Africa
Nicola Mulder
Mamana Mbiyavanga
Ziyaad Parker

Thomas Keane
Melanie Courtot
Jonathan Dursi

Salvador Capella
Dmitry Repchevski
JM Fernández

CINECA

DisGeNET
David Torrents

EU Can

Laura Furlong
Janet Piñero

Dean Hartley

Stephane Dyke

DNAStack
Marc Fiume
Miro Cupak

Phillip Struyf

Juha Törnroos
Teemu Kataja
Ilkka Lappalainen
Dylan Spalding

Beacon PRC
Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga
Alice Mann
Neerjah Skantharajah

"...and the others who shed their tears"
Team EGA - Barcelona
THANK YOU!

Core organizations:

Questions?

contact: babita.singh@crg.eu

Additional sources:

And infrastructure support from the following sources: