Data-driven Temporal Phenotyping of the GCAT Cohort from Electronic Health Records

Xavier Duran-Albareda

GCAT Genomes for Life

Institut de Recerca Germans Trias i Pujol (IGTP)
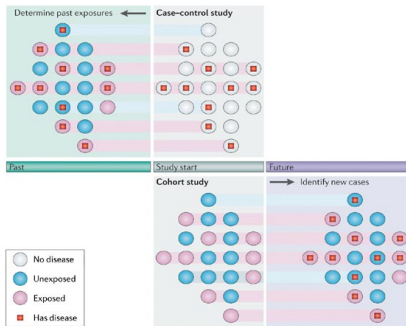
FI-2017 AGAUR

October 28th, 2016

A Cohort Study of the Genomes of Catalonia

- ▶ Genomic population-based cohort
- ▶ 25.000 participants by 2018
- ▶ General, asymptomatic population
- ▶ Analyze genomics and health interplay in the next 20 years

Cohort studies linked to Electronic Health Records

- Identify the **genomic basis** of **common disorders**
- **Electronic Health Records** are the major source of phenotype data available
- From hypothesis-driven to **data-driven** phenotyping
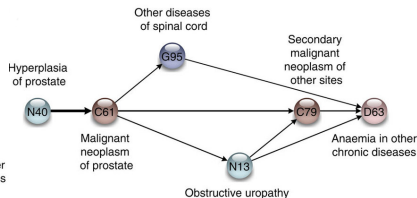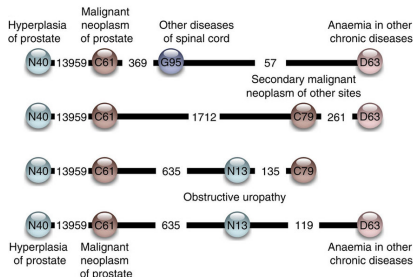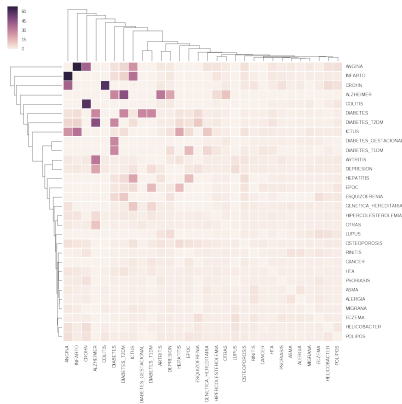
Temporal disease trajectories and disease clusters

Temporal disease trajectories condensed
from population-wide registry data covering
6.2 million patients

Anders Boeck Jensen[1,2], Pope L. Moseley[2,3], Tudor I. Oprea[1,3,4], Sabrina Gade Ellesøe[2], Robert Eriksson[1,2],
Henriette Schmock[5], Peter Bjødstrup Jensen[2], Lars Juhl Jensen[2] & Søren Brunak[1,2]
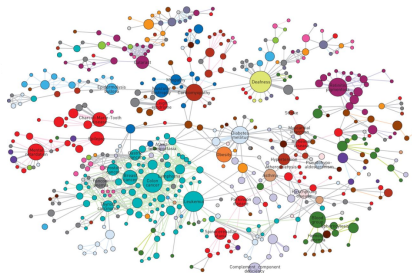
GCAT Snapshot and the Diseasome

Association rule mining
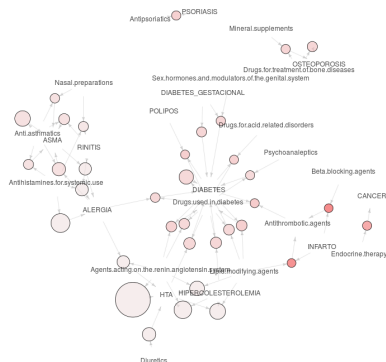
rules	support	confidence	lift	chiSquare
{Drugs.used.in.diabetes} => {DIABETES}	0.036209193	0.9357088	16.936870	3940.3809
{Agents.acting.on.the.renin.angiotensin.system} => {HTA}	0.156305245	0.3406028	3.958346	3690.2555
{Anti.asthmatics} => {ASMA}	0.042133176	0.8194842	10.971714	2963.0285
{HTA,Drugs.used.in.diabetes} => {DIABETES}	0.018709487	0.9338235	17.271919	2100.2424
{DIABETES,Lipid.modifying.agents} => {Drugs.used.in.diabetes}	0.014142605	0.8648649	22.493114	2084.5341
{ALERGIA,Anti.asthmatics} => {ASMA}	0.028874484	0.8376069	11.214350	2041.0416
{HIPERCOLESTEROLEMIA,Drugs.used.in.diabetes} => {DIABETES}	0.016352387	0.9652174	17.852577	1898.9645
{DIABETES,Agents.acting.on.the.renin.angiotensin.system} => {Drugs.used.in.diabetes}	0.013111373	0.8018018	20.852991	1778.5342
{Drugs.used.in.diabetes,Lipid.modifying.agents} => {DIABETES}	0.014142605	0.9142857	16.910549	1543.0976
{Agents.acting.on.the.renin.angiotensin.system,Drugs.used.in.diabetes} => {DIABETES}	0.013111373	0.8900000	16.461362	1386.7701
{Antihistamines.for.systemic.use} => {ALERGIA}	0.061431939	0.8424242	3.998864	1281.5520
{HIPERCOLESTEROLEMIA,Agents.acting.on.the.renin.angiotensin.system} => {HTA}	0.054065399	0.9683377	4.075063	1183.1037
{Drugs.for.treatment.of.bone.diseases} => {OSTEOPOROSIS}	0.010312316	0.8433735	17.723899	1174.1679
{DIABETES_GESTACIONAL} => {DIABETES}	0.009575722	1.0000000	18.495913	1148.2295
{RINITIS,Anti.asthmatics} => {ASMA}	0.014731880	0.8439655	11.848283	1091.6168
{Endocrine.therapy} => {CANCER}	0.004419564	0.9677419	36.698504	1074.8972
{HTA,Lipid.modifying.agents} => {HIPERCOLESTEROLEMIA}	0.043753683	0.8272981	3.525235	741.1850
{ALERGIA,Drugs.used.in.diabetes} => {DIABETES}	0.006629346	0.9375000	17.339918	737.7111
{Agents.acting.on.the.renin.angiotensin.system,Lipid.modifying.agents} => {HTA}	0.036535062	0.9117647	3.836986	710.8339

Diseases and medications rule network from GCAT cohort

Summary

- Data-driven temporal analysis of the cohort
- Identify subgroups of patients with distinct mechanisms of disease
- Genome associations with disease subgroups
- Risk models for common disorders
- Stratified medicine

Acknowledgements

Rafael de Cid, PhD

Ricard Gavaldà, PhD