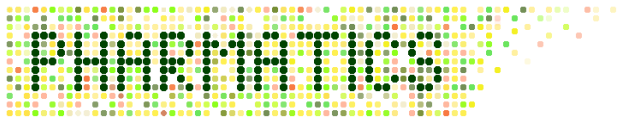
USING INFORMATION OF RELATED TRAITS TO IMPROVE GENETIC PREDICTIONS

Daniel Urda Muñoz

daniel@pharmaticsltd.com



21st May, Barcelona – Pharmatics Limited (Edinburgh, United Kingdom)

Think about your answer…

- Question: will we reach a point where genomic predictions may replace predictions based on rich clinical models?

Outline

1. Genotypic predictions: motivation
2. Including genomics into our models
3. Polygenic risk scores
4. Using information of related traits
5. Discussion

Outline

1. Genotypic predictions: motivation
2. Including genomics into our models
3. Polygenic risk scores
4. Using information of related traits
5. Discussion

Case of study

Prediction of response to an anti-rheumatic drug from genomics

1. Genotypic predictions: motivation

- Why?

1. Genotypic predictions: motivation

- Why?
  1. Use of genomics can improve **decision making**
  2. Use of genomics can improve our **understanding of disease**

1. Genotypic predictions: motivation

- Why?
  1. Use of genomics can improve **decision making**
  2. Use of genomics can improve our **understanding of disease**

- Pros and Cons:
  + Genotypes can be recorded from birth (or earlier)
  + In most cases, genotypes are almost the same through life

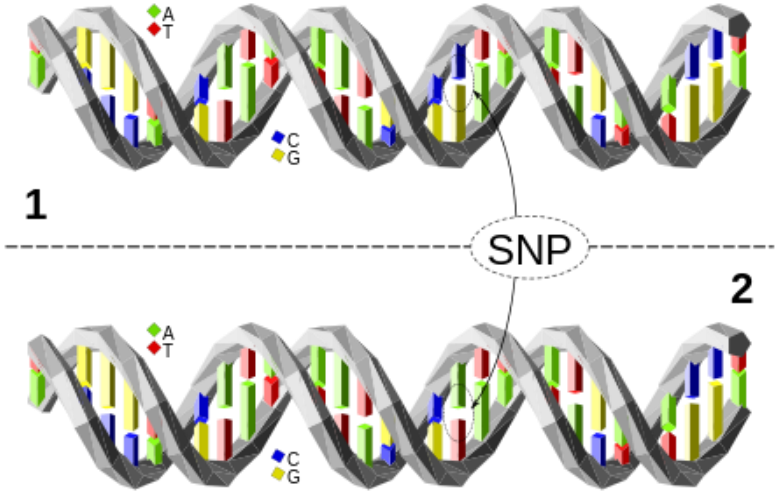
Prevention, diagnosis and treatment the soonest possible, previous to the appearance of any clinical symptom

  - Low prediction accuracy for most complex traits in humans

Trait variation depends not only in genetic but environmental factors

2. Including genomics into our models

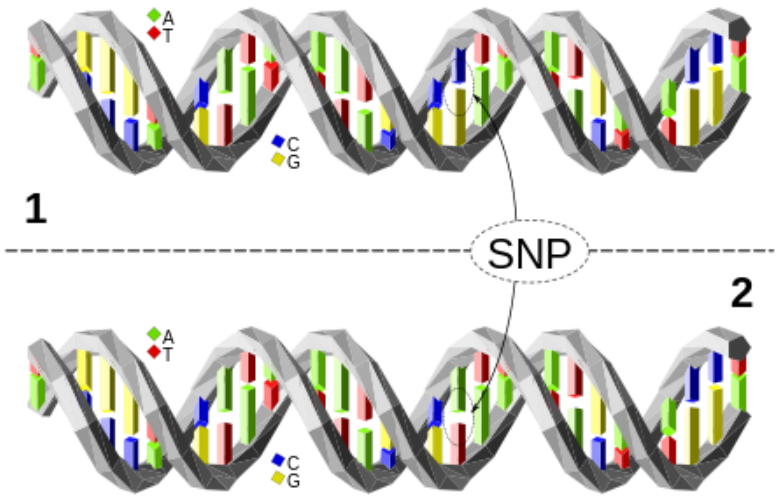
Single nucleotide polymorphism (SNP)



There are about 38 million SNPs in the human genome

2. Including genomics into our models

Single nucleotide polymorphism (SNP)



There are about 38 million SNPs in the human genome

Genome-Wide Association Studies

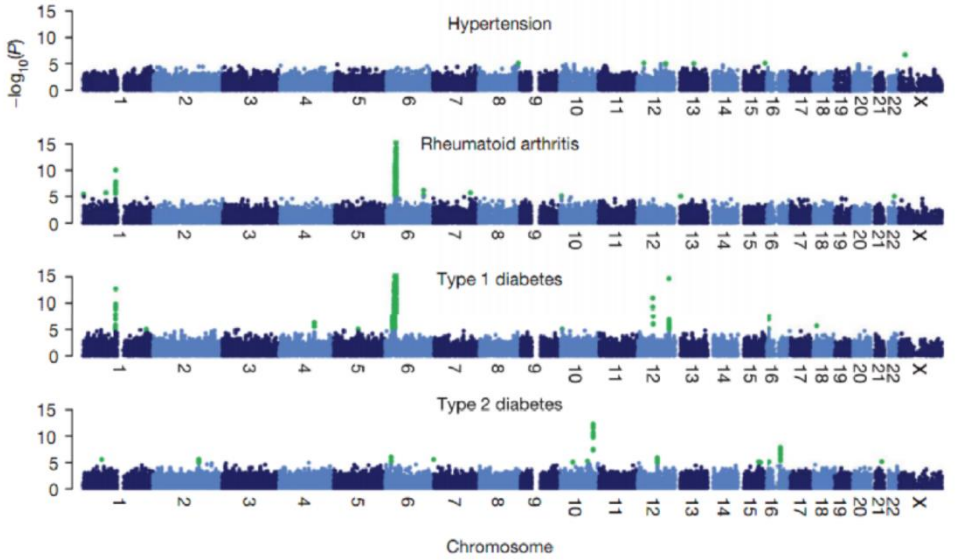
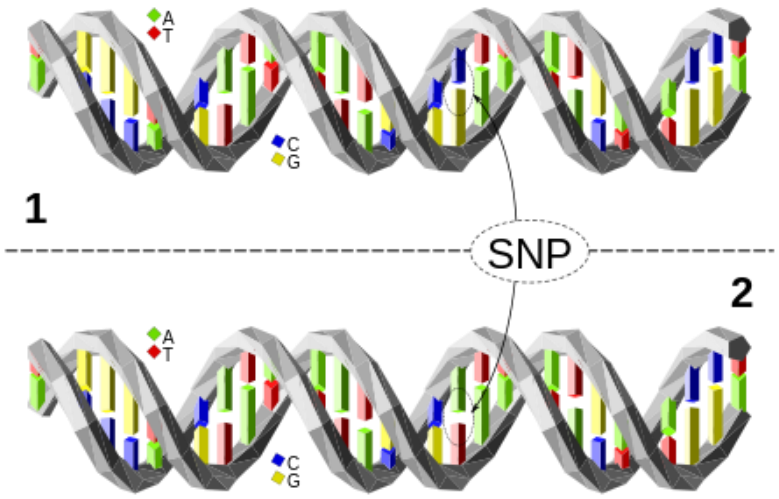


Figure: Genome-Wide Association Studies [WTCCC 2007, doi:10.1038/nature05911]

2. Including genomics into our models

Single nucleotide polymorphism (SNP)



There are **about 38 million SNPs in the human genome**

Genome-Wide Association Studies

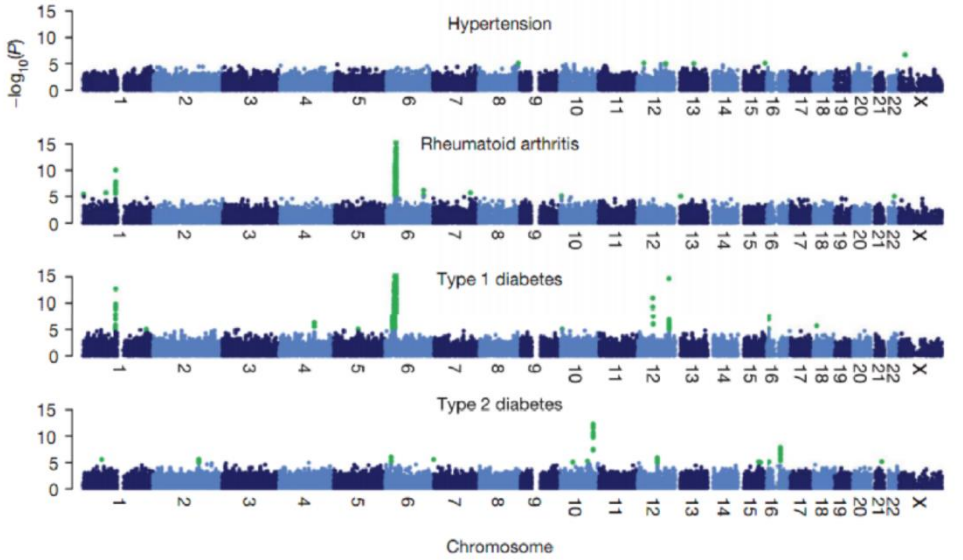


Figure: Genome-Wide Association Studies [WTCCC 2007, doi:10.1038/nature05911]

- How? Phenotype is modelled as a function of someone's genotype

  Example: harmful mutations at BRCA genes increase risk of breast cancer

  It can be modelled as a parametric function:

  (risk of breast cancer) = constant + B(BRCA mutations)*

2. Including genomics into our models

Option 1: include all SNPs using homogeneous priors

Many SNPs may not be important for our target trait
(Very) Large-p-Small-n scenario

2. Including genomics into our models

Option 1: include all SNPs using homogeneous priors

Many SNPs may not be important for our target trait
(Very) Large-p-Small-n scenario

Option 2: include only GWAS hits for our target trait

Usually, low proportion of the variance is explained

2. Including genomics into our models

Option 1: include all SNPs using homogeneous priors

Many SNPs may not be important for our target trait
(Very) Large-p-Small-n scenario

Option 2: include only GWAS hits for our target trait

Usually, low proportion of the variance is explained

Option 3: something intermediate between those two extremes

- Genome-Wide Association Meta Analysis (GWAMA)
- More generous p-values threshold (Bermingham et al. 2015, doi: 10.1038/srep10312)
- Use SNPs of related traits

2. Including genomics into our models

Option 1: include all SNPs using homogeneous priors

Many SNPs may not be important for our target trait
(Very) Large-p-Small-n scenario

Option 2: include only GWAS hits for our target trait

Usually, low proportion of the variance is explained

Option 3: something intermediate between those two extremes

- Genome-Wide Association Meta Analysis (GWAMA)
- More generous p-values threshold (Bermingham et al. 2015, doi: 10.1038/srep10312)
- Use SNPs of related traits

Overall, NOT combining prior knowledge with our data is a bad idea!

3. Polygenic risk scores

- One simple way (GWAS-based computation):
  1. Consider GWAS hits – those *p* SNPs associated with phenotype
  2. Multiply effect size by number of alleles at each locus
  3. Add-up across loci for each individual

$$PRS^{(i)} = \sum_{j=1}^{p} \beta_j x_j^{(i)}$$

GWAS Summary Statistics

LETTER

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height

A full list of authors and their affiliations appears at the end of the paper.

						STAGE 1 up to 133,653 samples			
SNP [a]	Chr	Position (bp)	Nearest/OMIM height gene [b]	Effect / other allele [c]	Frequency (effect allele)	Beta	*P*-value [d]	I^2	P_{het}
rs425277	1	2059032	***PRKCZ***	T/C	0.28	0.024	1.70E-06	0	0.73
rs2284746	1	17179262	***MFAP2***	C/G	0.48	-0.035	5.60E-15	17.77	0.07
rs1738475	1	23409478	***HTR1D***	C/G	0.59	0.022	1.90E-06	0	0.69
rs4601530	1	24916698	***CLIC4***	T/C	0.26	-0.024	2.00E-06	15.60	0.10
rs7532866	1	26614131	***LIN28***	A/G	0.67	0.022	3.30E-06	0	0.54
rs2154319	1	41518357	***SCMH1***	T/C	0.75	-0.034	4.30E-10	0	0.86

∗

Genotyped SNP Data

	A	B	C	D	E	F	G
1		rs425277	rs2284746	rs1738475	rs4601530	rs7532866	rs2154319
2	sample 1	0	1	2	2	0	0
3	sample 2	1	0	2	1	1	0
4	sample 3	0	0	2	2	0	1
5	sample 4	0	2	2	2	0	1
6	sample 5	1	1	1	2	0	0
7	sample 6	0	1	2	1	1	0
8	sample 7	1	0	2	1	1	1
9	sample 8	1	1	2	1	1	1
10	sample 9	0	1	1	2	1	0
11	sample 10	0	2	1	2	0	0

Figure: Constructing a polygenic risk score for height.

3. Polygenic risk scores

- One simple way (GWAS-based computation):
  1. Consider GWAS hits – those *p* SNPs associated with phenotype
  2. Multiply effect size by number of alleles at each locus
  3. Add-up across loci for each individual

$$PRS^{(i)} = \sum_{j=1}^{p} \beta_j x_j^{(i)}$$

- Advantages:
  - \+ Uses prior knowledge
  - \+ Privacy issues
  - \+ Dimensionality reduction

GWAS Summary Statistics

LETTER

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height

A full list of authors and their affiliations appears at the end of the paper.

						STAGE 1 up to 133,653 samples			
SNP [a]	Chr	Position (bp)	Nearest/OMIM height gene [b]	Effect / other allele [c]	Frequency (effect allele)	Beta	*P*-value [d]	I^2	P_{het}
rs425277	1	2059032	***PRKCZ***	T/C	0.28	0.024	1.70E-06	0	0.73
rs2284746	1	17179262	***MFAP2***	C/G	0.48	-0.035	5.60E-15	17.77	0.07
rs1738475	1	23409478	***HTR1D***	C/G	0.59	0.022	1.90E-06	0	0.69
rs4601530	1	24916698	***CLIC4***	T/C	0.26	-0.024	2.00E-06	15.60	0.10
rs7532866	1	26614131	***LIN28***	A/G	0.67	0.022	3.30E-06	0	0.54
rs2154319	1	41518357	***SCMH1***	T/C	0.75	-0.034	4.30E-10	0	0.86

*

Genotyped SNP Data

	A	B	C	D	E	F	G
1		rs425277	rs2284746	rs1738475	rs4601530	rs7532866	rs2154319
2	sample 1	0	1	2	2	0	0
3	sample 2	1	0	2	1	1	0
4	sample 3	0	0	2	2	0	1
5	sample 4	0	2	2	2	0	1
6	sample 5	1	1	1	2	0	0
7	sample 6	0	1	2	1	1	0
8	sample 7	1	0	2	1	1	1
9	sample 8	1	1	2	1	1	1
10	sample 9	0	1	1	2	1	0
11	sample 10	0	2	1	2	0	0

Figure: Constructing a polygenic risk score for height.

4. Using information of related traits

- Why don’t we use information about identified related traits to our target trait?

4. Using information of related traits

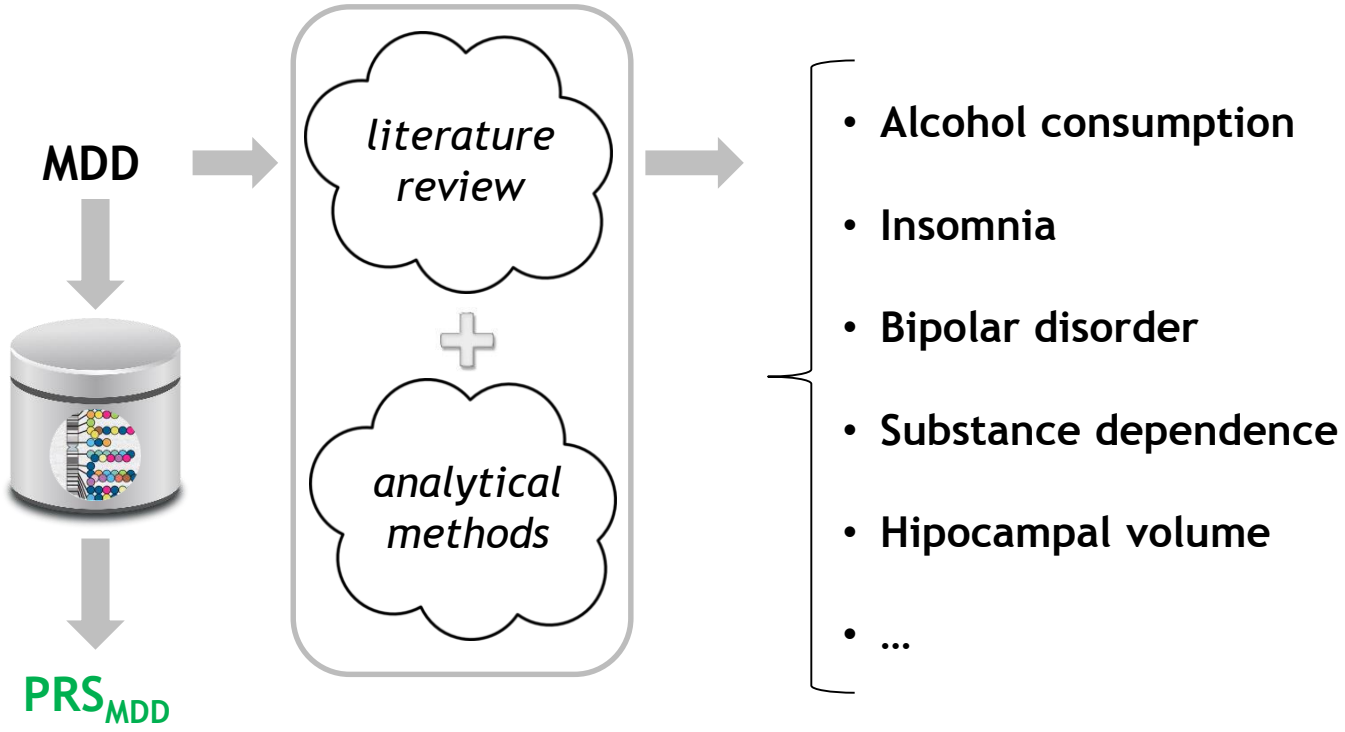
- Why don’t we use information about identified related traits to our target trait?

  Example: prediction of Major Depression Disorder (MDD)

4. Using information of related traits

- Why don't we use information about identified related traits to our target trait?

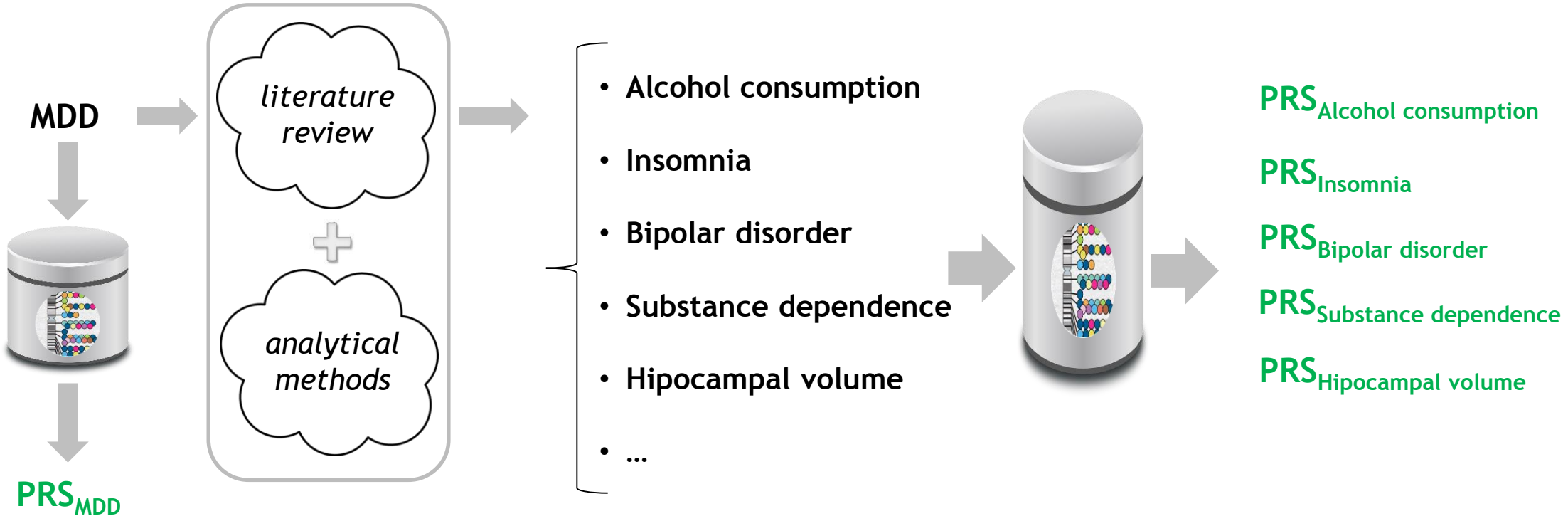
Example: prediction of Major Depression Disorder (MDD)



4. Using information of related traits

- Why don't we use information about identified related traits to our target trait?

Example: prediction of Major Depression Disorder (MDD)



4. Using information of related traits

- Results obtained in a real project (rheumatoid arthritis, RA):

Sample size: 304 individuals from a randomized clinical trial
Outcome: prediction of response to an anti-rheumatic drug
Models:

- Clinical, C1: about **45 clinical variables** forming a rich clinical model
- Genomics, G1: **172 PRS** (regional scores for RA and scores for gene expressions correlated with the RA regional scores)
- Genomics, G2: **642 PRS** for other related traits to RA

	Pearson σ (95% CI)
M0 (baseline)	0.53 (0.51, 0.55)
M1	0.59 (0.57, 0.61)
M2	0.59 (0.57, 0.61)
M3	0.56 (0.54, 0.59)

	Pearson σ (95% CI)
M0 (baseline)	-0.04 (-0.07, 0.01)
M1	0.05 (-0.01, 0.12)
M2	0.03 (-0.03, 0.08)
M3	0.05 (0.02, 0.10)

	Pearson σ (95% CI)
M0 (baseline)	0.12 (0.08, 0.17)
M1	0.16 (0.09, 0.23)
M2	0.16 (0.10, 0.24)
M3	0.16 (0.10, 0.24)

** Accuracy (correlation between predicted and observed phenotype) computed over the test samples by using 10-fold cross-validation repeated 20 times*

5. Discussion

- Question: will we reach a point where genomic predictions may replace predictions based on rich clinical models?
  - I will give my point of view later based on my personal experience among different projects

- Comments, ideas, suggestions…

Acknowledgements

I am grateful for financial support from the European Union 7th Framework Programme through the Marie Curie Initial Training Network "Machine Learning for Personalized Medicine" MLPM2012, Grant No. 316861.

- Felix Agakov
- Athina Spiliopoulou
- Paul McKeigue
- Marco Colombo

Many thanks!