Big Data approaches for the estimation of Biological Age: a new perspective for the Moli-sani Study

Alessandro Gialluisi, PhD

Department of Epidemiology and Prevention,
IRCCS Istituto Neurologico Mediterraneo Neuromed, Pozzilli (IS)

Resident in Medical Statistics and Biometry, Sapienza University of Rome

alessandro.gialluisi@moli-sani.org
The Moli-sani Study is a population-based, prospective cohort study including **24,325 citizens** (men and women, aged ≥35 years) of the Molise region (enrolled March 2005-April 2010), with the purpose of investigating genetic and environmental risk factors for many complex disorders.
The Moli-sani study

Main sources of data available

- personal and family history of disease
- health-related behaviors (e.g. smoking habits, physical activity)
- dietary patterns (based on food frequency questionnaires)
- socioeconomic status (e.g. educational attainment)
- psychometric scores (e.g. anxiety and depressive symptoms)
- instrumental variables (ECG, spirometry, blood pressure)
- anthropometric measures (e.g. height, weight, waist-to-hip ratio)
- comprehensive biochemistry and hemochromes blood test
- selected metabolites in urine (microalbumin, ongoing)
Moreover, the study allows:

- access to medical registries (**hospitalizations**)
- access to demographic records (**mortality** registry)
- linkage with **drugs prescription** database
- geolocalization and linkage with geodatabases (e.g. PM$_{2.5}$ and PM$_{10}$ maps)
- cohort recall (**active follow-up**, ongoing)
Big Data in the Moli-sani study

<table>
<thead>
<tr>
<th>Source</th>
<th>Variables#</th>
<th>Observations#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet Questionnaires</td>
<td>1,600</td>
<td>38,920,000</td>
</tr>
<tr>
<td>Spirometry</td>
<td>153</td>
<td>3,721,725</td>
</tr>
<tr>
<td>ECG</td>
<td>617</td>
<td>15,008,525</td>
</tr>
<tr>
<td>Clinical history</td>
<td>2,100</td>
<td>51,082,500</td>
</tr>
<tr>
<td>Family history of disease</td>
<td>841</td>
<td>20,457,325</td>
</tr>
<tr>
<td>Circulating biomarkers</td>
<td>592</td>
<td>14,400,400</td>
</tr>
<tr>
<td>Follow-up data</td>
<td>680</td>
<td>16,564,664</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6,583</strong></td>
<td><strong>160,155,139</strong></td>
</tr>
</tbody>
</table>

How can we be part of the Big Data revolution?
A starting point..

Estimation of Biological Age
(BA, i.e. the hypothetical underlying age of an organism):

• How to estimate it through Machine Learning (ML) approaches?
  • How does this differ from Chronological Age (CA)?
• Is there any relation between $\Delta age = BA - CA$ and measures of interest in public health (e.g. frailty, mortality)?
  • Can we possibly exploit $\Delta age$ in personalized medicine?
• Can we estimate an organ- or system-specific BA and $\Delta age$?
Estimation of BA from blood biomarkers: a foundation work

Deep biomarkers of human aging: Application of deep neural networks to biomarker development

Evgeny Putin 1,2, Polina Mamoshina 1,3, Alexander Aliper 1, Mikhail Korzinkin 1, Alexey Moskalev 1,4, Alexey Kolosov 5, Alexander Ostrovskiy 5, Charles Cantor 6, Jan Vijg 7, and Alex Zhavoronkov 1,3

1 Pharma.AI Department, Insilico Medicine, Inc, Baltimore, MD 21218, USA
2 Computer Technologies Lab, ITMO University, St. Petersburg 197101, Russia
3 The Biogerontology Research Foundation, Oxford, UK
4 School of Systems Biology, George Mason University (GMU), Fairfax, VA 22030, USA
5 Invitro Laboratory, Ltd, Moscow 125047, Russia
6 Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
7 Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

ML approach for the estimation of BA

Pipeline:

1. Blood biomarkers preprocessing (normalization and outliers detection)
2. Training and testing of the Deep Neural Network (DNN) models (90:10)
3. Ensembling of 21 DNNs providing high predictive value in BA estimation

Putin et al., 2016
The best ensemble model (21 DNNs) showed $R^2=0.83$ and ε-prediction accuracy = 83.5% within the 10 year frame ($\varepsilon = 10$)
DNN outperformed six other ML techniques: GBM (Gradient Boosting Machine), RF (Random Forests), DT (Decision Trees), LR (Linear Regression), kNN (k-Nearest Neighbors), ElasticNet.

Putin et al., 2016
Is BA estimation population-specific?

Is BA estimation population-specific?

Mamoshina et al., 2018
<table>
<thead>
<tr>
<th>Testing Set</th>
<th>Training Set</th>
<th>Features</th>
<th>Pearson’s r*</th>
<th>$R^2*$</th>
<th>MAE (years)*</th>
<th>$\epsilon$-Accuracy*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canada</strong></td>
<td>Canada</td>
<td>20</td>
<td>0.70 [0.70; 0.70]</td>
<td>0.52 [0.52; 0.47]</td>
<td>6.36 [6.46; 6.28]</td>
<td>0.80 [0.79; 0.81]</td>
</tr>
<tr>
<td></td>
<td>S. Korea</td>
<td>20</td>
<td>0.57 [0.55; 0.59]</td>
<td>0.24 [0.24; 0.24]</td>
<td>7.87 [7.66; 8.12]</td>
<td>0.70 [0.73; 0.69]</td>
</tr>
<tr>
<td></td>
<td>E. Europe</td>
<td>20</td>
<td>0.52 [0.52; 0.52]</td>
<td>0.24 [0.27; 0.22]</td>
<td>9.68 [9.98; 9.2]</td>
<td>0.62 [0.58; 0.70]</td>
</tr>
<tr>
<td><strong>S. Korea</strong></td>
<td>Canada</td>
<td>20</td>
<td>0.52 [0.52; 0.47]</td>
<td>0.24 [0.27; 0.22]</td>
<td>7.1 [7.66; 8.12]</td>
<td>0.72 [0.72; 0.71]</td>
</tr>
<tr>
<td></td>
<td>S. Korea</td>
<td>20</td>
<td>0.70 [0.73; 0.66]</td>
<td>0.49 [0.53; 0.43]</td>
<td>5.59 [5.45; 5.77]</td>
<td>0.85 [0.86; 0.84]</td>
</tr>
<tr>
<td></td>
<td>E. Europe</td>
<td>20</td>
<td>0.54 [0.56; 0.52]</td>
<td>0.29 [0.27; 0.31]</td>
<td>9.77 [10.21; 9.02]</td>
<td>0.65 [0.62; 0.68]</td>
</tr>
<tr>
<td><strong>E. Europe</strong></td>
<td>Canada</td>
<td>20</td>
<td>0.68 [0.68; 0.72]</td>
<td>0.27 [0.24; 0.27]</td>
<td>9.25 [10.09; 7.91]</td>
<td>0.38 [0.37; 0.38]</td>
</tr>
<tr>
<td></td>
<td>S. Korea</td>
<td>20</td>
<td>0.76 [0.77; 0.75]</td>
<td>0.34 [0.36; 0.33]</td>
<td>8.52 [8.89; 7.93]</td>
<td>0.31 [0.28; 0.35]</td>
</tr>
<tr>
<td></td>
<td>E. Europe</td>
<td>20</td>
<td>0.84 [0.85; 0.82]</td>
<td>0.69 [0.72; 0.67]</td>
<td>6.25 [6.24; 6.28]</td>
<td>0.82 [0.83; 0.80]</td>
</tr>
</tbody>
</table>

* Accuracy stats are reported in the following format: all [females; males]
Most predictive markers in BA
• **Hypoalbuminemia** associated with several pathophysiological conditions and ageing
  • **Glucose** levels tend to increase with age
    • Age-related decrease in **haemoglobin** is common in the elderly
  • **Urea** levels also increase with age (due to age-related decrease in muscle mass)
Δage predicts mortality risk

In the general population subjects predicted younger than their CA (Δage < -5) have a lower mortality risk, while subjects predicted older (Δage > 5) have a higher risk than subjects with BA ~ CA (-5 < Δage < 5)

Mamoshina et al., 2018
Organ-specific BA estimation

Women with high dietary intake of antioxidants showed increased forced expiratory volume in the first second (FEV$_1$) and forced vital capacity (FVC), corresponding to a pulmonary age improvement of $\sim$3.3 years.

Pulmonary age can be estimated based on spirometry measures (forced expiratory volume and forced vital capacity)
Brain Age (BrA) can be estimated based on structural neuroimaging features.

Brain-specific BA estimation

Brain Age (BrA):

- provides very accurate estimate of CA (Pearson’s $r \geq 0.94$, $R^2 \geq 0.91$ and MAE $\sim 4$ years)
- based on Gaussian Processes Regression (GPR)
- best performance when both matter (GM) and white matter (WM) features are included in the model
- high levels of within-scanner reliability (Intra-Class Correlation $= 0.94$) and good levels of between-scanner reliability (ICC $= 0.66$)
- is moderately heritable ($h^2 \geq 0.5$)
Brain-specific BA (BrA) and health

$\Delta_{\text{BRAINage}} \text{ (BrA - CA)}$: 

- increased in Alzheimer Disease, focal epilepsy, Down Syndrome, HIV and traumatic brain injury 

- negatively associated with years of education and self-reported physical activity 

- associated with cognitive performance 

- associated with measures of frailty (weaker grip strength, poorer lung function, slower walking speed) 

- predicts mortality risk in the general population 

Cole & Franke, 2017
Future perspectives for the Moli-sani study (1)

Build a BA estimation model for the Italian population:

• based on a high number of circulating biomarkers (blood-based BA)
• integrating additional (instrumental, anthropometric) variables (“holistic BA“)
• test the predictive capacity of health events occurrence (health conditions and mortality) by Δage
• test the relation of Δage with environmental factors
  ➢ Lifestyle (e.g. diet, physical activity)
  ➢ Pollution (e.g. PM exposure maps)
Develop organ- or system-specific approaches for BA estimation (e.g. brain):

- improve accuracy through method development (e.g. including connectivity measures?)
- how is $\Delta_{\text{BRAINage}}$ influenced by environmental (e.g. lifestyle) factors?
- investigate the genetic basis of $\Delta_{\text{BRAINage}}$
- test its relation with prevalent health conditions (e.g. cardiovascular disease, diabetes, obesity)
- test its efficacy in predicting clinical events risk
Overall final goal

Test efficacy and applicability of Δage in epidemiology:

• can we build risk maps for Δage? (at the organism, system and organ level)

• how to reduce Δage? (Δage-oriented public health policies)

A total shift in the paradigm of public health policies, focused on healthy ageing rather than on ageing-related diseases
Thanks for your attention!