Ann Ist Super Sanità 2018 | Vol. 54, No. 3: 174-175
DOI: 10.4415/ANN_18_03_02

COMMENTARY

COMMENTARY
Big Data. A briefing

Virginia Todde and Alessandro Giuliani

Dipartimento Ambiente e Salute, Istituto Superiore di Sanità, Rome, Italy

Abstract
The data deluge (generally referred as "Big Data") biomedical scientists are facing in these years asks for a serious epistemological thinking in order to avoid both "data bases idolatry" and "preconceived refusal". Starting from the evident reproducibility crisis of biomedical sciences here we sketch some hopefully useful indications for a sensible use of data mining approaches.

Key words
- Big Data
- biomedical data

SETTING THE FRAME

The term "Big Data" encompasses all the data collections endowed with a sufficient "size" and lack of definition (having been assembled with no *a priori* hypothesis or specific research task) to be considered as still largely unspoiled territories from where to derive new insights in the form of unforeseen regularities. In analogy with the exploration of the frontier lands during the gold rush, the exploration of such huge data repositories is usually referred to as "Data Mining". In both biology and medicine literature such approaches are exponentially rising in frequency: virtually any *in vitro* or *in vivo* approach is complemented by an *in silico* section, where experimental results are validated by the comparison with results emerging from big data sets. This way of doing blurs the traditional concepts of "scientific evidence" and, at least in our opinion, is provoking a substantial reshaping of biomedical research.

STATE OF THE AFFAIRS

Scientists tend to consider statistical methods as a largely "content independent" set of tools to be used in the "hypothesis testing" final phase of their work. This implies a neat separation of the "hypothesis generating" and "critical evaluation" of the experiment from the statistical testing phase.

Typically a biologist decides to focus on some descriptors (gene expression levels, protein concentrations, cell counts etc.) considered as instrumental for elucidating a given problem of interest, on the basis of his/her specific knowledge of the studied phenomenon (*hypothesis generating phase*).

The obtained results are then collected and analysed in the context of few standard statistical frames: 1) presence of significant correlations among a set of descriptors; 2) presence of a relevant effect exerted by a control parameter (i.e. a drug administered at different doses) on a variable of interest; 3) discovery of a superposition between classifications of the same objects based on different metrics (i.e. concordance between genetic distances and phenotypes of a set of organisms or cell populations) (*hypothesis testing phase*).

The congruence of the results coming from hypothesis testing phase with the starting hypotheses is critically evaluated in the context of existing literature (*critical evaluation*).

The above sketched process implies an ancillary role of statistical methodology that in some cases is demanded by a professional statistician whose role is suggesting the "best way" to highlight what the "core biomedical team" wants to know, but the two "semantically relevant" parts of the study (the first and the third) do not really interact with the methodological issues (confined in the second).

Such "classical" way of managing biomedical research faced a "reproducibility catastrophe" in these last ten years [1-3].

The essence of this crisis is related to "overfitting": in presence of too many degrees of freedom (being they different variables to be analysed, different experimental conditions..) as consequence of the development of 'high-throughput' techniques allowing to measure thousands of different descriptors of (relatively few) independent observations, the risk of chance correlations becomes too high [4].

To face this crisis that is appearing as a fatal menace for knowledge advancement, one of the response was (roughly). "Let's give up with contingent (and largely biased) theory-driven experimentations looking at this or that specific mechanism and look, without preconceived ideas, to the "whole-thing" (the development of various -omics makes this possible): the emerging correlations will allow for new ideas and findings spontaneously appear in a data-driven way".

This is the attitude set forth by "Big Data" diffusion that pushes biologists to actively face the rapidly chang-

Address for correspondence: Alessandro Giuliani, Dipartimento Ambiente e Salute, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy. E-mail: alessandro.giuliani@iss.it.

ing way of doing science.

It is sufficiently clear that the pure enumeration of single relevant correlations across a huge number of variables only exacerbates the reproducibility crisis, making "pure data-driven" approaches proposed by the "Big Data" extremists claiming for the "end of scientific method", (see for example [5] the Heaven of chance correlations (see [6] for a very interesting critics to the purely informatics approach to science).

Other sciences, like physics and chemistry have since long time recognised in any scientific modelling effort, the simultaneous presence of "sloppy" (contingent, noisy, idyosincratic) and "stiff" (stable, reliable, generalizable) features [7]: a good piece of science must discriminate between "stiff" and "sloppy" part of information.

How this discrimination is possible when facing "Big Data" in biology, where we cannot rely (like happens in physics and chemistry) on reliable quantitative theories that allow a strict *a priori* filtering of incoming information?

SENSIBLE USE OF BIG DATA EXPLORATIVE APPROACHES

The scientist does not know in advance if a given variable will be of use for his/her goal or not, thus it is a good rule to consider more observables than necessary (this is why "Big Data" are potentially very useful in biomedical sciences), leaving the subsequent steps of the analysis to "keep alive" only the relevant part of information.

Jean Paul Guigou [8] has aptly defined the multidimensional methods as those that allow the simultaneous treatment of variables "which are numerous, approximate, not very significant (in the sense that each of them, singly, carries limited information), discrete or continuous, heterogeneous, qualitative or quantitative". This is exactly the situation of a data mining enterprise: the plethora of "approximate, not very significant" pieces of information must collapse into few relevant "coarse grain" integrated descriptors emerging as "meaningful summaries" that keep only the "correlated part of information", i.e. the consensus among many different descriptors condensed into global scores (principal component analysis is the by far most common tool to obtain this goal, and stands as a perfect example of systemic attitude since more than one century [9]).

This implies "the whole" is not an unstructured collection of pieces of information but an organized system of relations among variables. This systemic view asks for a direct engagement of biologists (experts-in-the-field according to data science jargon) into statistical analysis.

The distillation of shared (and thus robust) information initially dispersed into a plethora of descriptors can by no means intended as a pure statistical procedure devoid of any link with the specific biological question. On the contrary, the progressive emerging of a relation structure from the raw data set corresponds to what we called "hypothesis generating" phase and asks for a "critical evaluation" by the biologist that must critically follow the entire data analysis procedure that is no more a largely ancillary (and thus prone to be delegated) hypothesis testing procedure but the actual scientific work. The conflation of the three classically distinct phases of "hypothesis generating", "hypothesis testing" and "critical evaluation" into a single organic activity asks for a reshaping of the scientific culture and education toward a recovery of the "generalist" scientist (that is practically extinct since more than fifty years). The end of over-specialization could be the "sun" shining at the end of the information crisis tunnel.

Conflict of interest statement

None to declare.

Accepted on 28 June 2018.

REFERENCES

1. Ioannidis, John PA. Why most published research findings are false. *PLoS Medicine* 2.8 2005;e124.
2. Joyner MJ, Paneth N, Ioannidis JP. What happens when underperforming big ideas in research become entrenched? JAMA. 2016;316(13):1355-6.
3. Thiago FA, Monserrat JM. Reproducibility crisis in science or unrealistic expectations? EMBO Reports. 2018;e46008.
4. Young SS, Karr A. Deming, data and observational studies A process out of control and needing fixing. Significance. 2011;8(3):116-20.
5. Anderson C. The end of theory. The data deluge makes the scientific method obsolete. Wired magazine. 2008;16(7):16-7.
6. Calude CS, Longo G. The deluge of spurious correlations in big data. Foundations of science. 2017;22(3):595-612.
7. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective. Sloppiness and emergent theories in physics, biology, and beyond. J Chemical Phys. 2015;143(1):07B201_1.
8. Guigou JP. Methods multidimensionelles. Paris: Dunod; 1977.
9. Giuliani A. The application of principal component analysis to drug discovery and biomedical data. Drug discovery today. 2017;22(7):1069-76.