EXPOsOMICS from a Big Data perspective: Interview with Professor Paolo Vineis, PI of the EXPOsOMICS Project

As the EXPOsOMICS project is entering its 3rd year. Prof Paolo Vineis talks to Dr Stefano Canali about the project and takes a fresh look at the EXPOsOMCS project through the perspective of philosophy of science.

Interviewer: Stefano Canali (SC)
Interviewee: Prof. Paolo Vineis (PV)

SC: The EXPOsOMICS project studies the relation between exposure to environmental factors and chronic disease. Through innovative statistic methods, it aims at individuating biomarkers that are capable of tracing the exposure to environmental factors and the development of the disease. The objective is to find associations between these two, i.e. biomarkers of exposure and biomarkers of disease. Thanks to these biomarkers one can make predictions. For instance, if we know that there's been an exposure to a certain type of element of the air; we might be able to say that there is an increased probability of getting a disease, in particular if relevant biomarkers are detected. Besides predictions, your goal is also to influence policies and suggest a series of environmental and public health policies. Can you tell me more?

PV: Yes the project has also a policy aspect. For example, I am currently working at a document for the US National Academy of Sciences on how OMICS methods can be used to predict disease and possibly substitute for instance animal tests in the future, if we become able to identify robust molecular alterations that consistently predict disease. One of the goals of the committee of the US National Academy of Sciences is to be able to precisely estimate risks associated with different levels of exposure using biomarkers as predictors.

SC: How do you study the statistical associations between biomarkers of exposure and biomarkers of disease?

PV: There are two steps to the study of associations: one is agnostic, meaning that associations are searched without a priori hypotheses; the other, instead, is guided by a priori hypotheses, for instance biologically-informed (e.g. through prior experimental knowledge) biological pathways affected by a certain exposure and/or involved in the disease development. For example, oxidative damage to DNA is a typically investigated pathway.

SC: An important aspect of your work is the meet-in-the-middle approach. From a certain point of view, with this approach you go beyond statistical associations. When a statistical association is found in the data, the aim of the approach is validating the association using biomarkers, because there may be issues with a complete reliance on statistical associations, such as confounding.

PV: Yes, sure. There are two things worth mentioning here. One is Bradford Hill's guidelines. Bradford Hill was an English statistician who dealt with the issue of tobacco smoking when tobacco companies denied that there was a relation with lung cancer, saying that it was only a statistical association and not a causal relationship. He established a set of guidelines allowing to establish when it is probable that the relation is causal and not only statistical; one of these guidelines is biological plausibility, which means that for example there are changes in intermediate markers or molecules like DNA that reinforce the

causal nature of the observation. So Bradford Hill's is one approach. The other approach I refer to is a philosophy of science text that is a reference book for me, i.e. Wesley Salmon's book that introduced the idea of "propagation of a mark", which is very similar to the meeting-in-the-middle approach (the propagation of a mark from exposure to disease). This is clear in the case of smoking, because we can measure for instance the metabolites of nicotine; then we can measure the aromatic hydrocarbons binding with haemoglobin and albumin, then the same chemicals binding with DNA; then we can identify DNA mutations in smokers comparing them with non-smokers; then we can identify alterations of a functional type in DNA such as epigenetics. Finally we can link all these events with the same or similar events found in lung cancer cells (specifically certain gene mutations like P53). So yes, the inspiration for the meet-in-the-middle comes from Wesley Salmon. In practice the meet-in-the-middle is extremely simple as it relies on the identification of biomarkers that are both reflecting effects of exposures and also contributing to future disease risk.

SC: The search for causality is necessary for your research. Isn't it?

PV: Yes and the approach is also important in order to go beyond the limits of purely descriptive epidemiology, where some exposure is linked with disease, but actually not much can be said about biological plausibility and there may be confounding, etc.

SC: Regarding agnostic research, many, including you in one of your articles, speak of 'data-driven' research. According to this idea, one does not have to elaborate theories, hypothesis, and models and then test them against the data; it is the data which guides research, without the need for any theory, hypothesis, model, etc. Can this really be done without existing knowledge?

PV: Actually there has been quite a radical change driven by technology, so research is 'technology-driven' besides 'data-driven'; this has taken place with GWAS that is the studies where researchers look at hundreds of thousands of genetic variants (single nucleotide polymorphisms). In these studies there is literally no a priori hypothesis, in the sense that all the genetic variants are investigated and prior knowledge is not used to filter out some a priori irrelevant variants: so everything is a posteriori. However, the interpretation of data is clearly done on the basis of existing knowledge; for instance if you see that the variant of a certain gene influences the risk of heart arrhythmia, you look at the function of the gene and whether the association has biological plausibility (considering that the confounding in the case of genetic associations is less problematic than with environmental exposures). There has been a shift because epidemiology has always insisted, if you look for example at textbooks, on the formulation of a priori hypotheses, and the importance of the study design which should be guided by a priori hypotheses. To reduce computational burden and ensure both feasibility and interpretability of statistical analyses, prior knowledge is not fully discarded and Bayesian approaches are typically aiming at that. This hypothesis-driven philosophy has been largely abandoned with omics epidemiology, which places research in an exploratory context, seeking for the identification of novel, unreported findings from highly complex and large data: that research becomes somehow technology-driven and can be viewed as a fishing exercise. Now much research is done with metabolomics, for instance, and with epigenetics: you generate a large amount of data and then a posteriori you look at their meaning. While these investigations do not necessarily call upon prior knowledge, it is interesting to note that the understanding, and ultimately the validation of these novel results cannot be achieved without integrating prior knowledge to validate their biological relevance.

***SC:* Can we say that the knowledge you produce from the data is a consequence of an integration of a series of statistical instruments?**

PV: Yes, I'd first say that different kinds of data require different statistical approaches, depending for example on the type, size, and complexity of the data. There is also a sort of hierarchy of approaches for a given type of data: as an exploratory approach first we begin with the univariate analysis, then we check whether in the data there is evidence for potential joined effect: sets of omic signals being more strongly associated to exposures or disease risk than the sum of the association of each signal separately. In the latter case (potentially identified by the existence of correlations), we use multivariate models, which include dimension reduction techniques: for instance PCA [principal components analysis], which builds upon the correlation across omic signals to summarise the data without losing information making data more tractable and defining homogeneous pools of informative signals.

***SC:* So, the complexity of both the target systems and the data requires a plurality of tools.**

PV: Yes. I'd say that statistical models have been developed mainly as a consequence of two needs, methodological and biological. Methodological needs: looking at certain data structures mainly driven by correlation structures within the data which needs to be modelled to inform the redundant information carried across omic signals. Biological needs are different, for instance in the case of cancer we know that there is not only one cause (a necessary cause). We know only one risk factor which is necessary for cancer onset: HPV for cancer of the cervix uteri. Otherwise, we know no necessary cause of cancer and it is probable that single cases of cancer are due to a variety of exposures, even weak exposures; for instance atmospheric pollution is a known risk factor with strong evidence of carcinogenicity, but the association with cancer is weak and probably it involves complex interactions with other risk factors. The relationship between exposure, risk factors and disease is complex, and requires refined models (e.g. causal network models) which are able to assess causality: typically adding to the establishment of a link between markers and either exposures or disease risk, the assessment of a direction in that link (which marker causes what).

***SC:* Recently there have been many works in philosophy of science looking at what it means to curate the data and at the kind of data curation carried out in science. I've seen that in a few cases you don't directly collect data, but you use already existing databases; if, instead, you have to collect data, do you carry out a specific kind of curation?**

PV: Well, we do both things. The typical example may be metabolomics, where on the one hand (for instance) blood samples are chemically analysed through mass spectrometry or nuclear magnetic resonance, and thousands of signals are derived. The first step is a sort of data curation and is called pre-processing. It means that we look at out-of-range data points, which are clearly outliers; then, in the case of mass spectrometry we remove the peaks which are relative to molecules which we are not interested in but are abundant; then we look at those variables that influence the quality of data, called nuisance parameters. Actually there is a whole series of pre-processing steps. For instance, metabolomics or other techniques may be influenced by the day in which the

measurement has been carried out, or the operator who has used the machine, or the reagents (the batch effects). This is quite evident for epigenetics: let's say that we have 500 samples but we can analyse one hundred of them at a time (for technical constraints); therefore we divide them into five batches, and there is almost always some degree of discrepancy between a batch and another, that is if we use the same sample for five different batches we have five slightly different results. Pre-processing is a way to curate the data in order to adjust for example for the batch effect, one of the nuisance variables. There are some solutions to calibrate the data across batches, including the use of quality control samples that are run in each batch. When pre-processing the data, measurements are then slightly transformed to ensure that the results from the quality control samples (which are the same sample across batched) in consistently measured across batches. Further steps are related to the interpretation of the findings. For example, in metabolomics we use databases which are key to annotation (i.e. identify which signal corresponds to which molecule). This consists in interpreting the spectrometry peaks: for instance we find 10 peaks which are statistically associated with colon cancer in our observations; those 10 are usually the ones which have withstood all the statistical tests (including correction for multiple comparisons) and the pre-processing procedures. At this point we try to understand what those 10 peaks mean in chemical terms (mass spectrometry only tells us which are the peaks, not which molecules they are). In order to discover what these molecules are, we can look at databases and this is called annotation. If we are lucky enough, some of those peaks can be found in the databases (based on their mass and other chemico-physical properties like retention time); usually this is not the case, so annotation requires other chemical work to identify precise molecules.

SC: ***You were saying that there's been a great change in the technology-driven sense: the change consists, on the one hand, in the possibility of carrying out agnostic research and, on the one other hand, in working on a huge quantity of data.***

PV: Yes, there has been a huge increase in what is called throughput that is the possibility of quickly analysing thousands of samples. For instance with GWAS, 2,000/3,000 samples for millions DNA variants can be analysed in a few weeks. So there has been this increase in the analytical power, but while tremendous improvements in the computational power have been achieved in the past decade, in-depth statistical analyses, and more importantly, biological interpretation and full exploitation of the results still remain suboptimal.

Stefano Canali is a Master of Science student at the Science and Technology Studies Department of UCL, London. He is interested in studying science and technology, both at the theoretical and practical level; in particular, his main interested is in digital technologies and the ways these are changing scientific practice and epistemology. His Master's degree is on the epistemology of big data and causality.