Precision medicine is an emerging field whose relevance is destined to grow at a pace, that at least in part, depends on the concomitant development of Electronic Health Records. Digital medicine has set the organization of the latter as its priority in an attempt to find synergies between phenotypes and markers from diseases and comorbidities. We present a rationale for motivating roles and functions of systems-wise multi-evidenced and algorithmically agnostic frameworks whose expected deliverables are: a) Personalized decisions and b) Risk profiles at stratified scales (individual to populations across communities).

For an effective interoperability, a few synergistic elements must be optimized: data re-use and novel evidences, predictive models for assessment, algorithm efficiency and portability, development of disease ontologies and of multiplexed networks to handle seamlessly the complexity of their relationships.

**Box 1. Electronic Health Systems (EHS): advantages and limitations.**

**EHS advantages**
- Electronic Health Systems allow EHR synergies aimed at deep phenotyping and marker re-modulation.
- Clinical decision support systems (CDSS) are valuable deliverables.
- Strong impacts on complex diseases and comorbidities are expected.
- Interdependent networks and community structures may be inference drivers.

**EHS limitations**
- Geo-differentiation and ethnic balance critical to widespread adoption of EHRs.
- Protocols for sharing digital information needed for effective aggregation of patients information.
- Records distillation for sufficient and accurate information to make decisions, still facing incompleteness of data and diversity of digital forms.
- Embedding of CDSS crucial for standardization of records and actionable decision making.
Electronic Health Systems

Electronic Health Records (EHR)³ require interoperability at systems scale to optimize the chances of finding patterns linked to both causal and correlative dynamics. This calls for a definition of the role of precision medicine³ in the presence of a variety of experimental evidences and a multitude of data types, and given the implications for personalized and population health management.

Behind co-occurring patterns correlated with, or causally relevant for clinical assessment, there are biological, physical, behavioral, environmental and technological processes. These act as Big Data Generators of heterogeneous information transferred by deterministic and stochastic signals at various spatiotemporal scales. With additional contextual data, a strong impact for quality care can be achieved.

An Electronic Health Systems (EHS) (Figure 1) allows for Big Data management within a common environment. The key component of methods enables inference through the selection of features and the extraction of statistics from active learning algorithms and automatic decision rules. The synergism between data and methods, once contextualized, is channeled towards: 1. Personalized (diagnostic, therapeutic, preventive) decisions and 2. Risk profiles at both individual and composite (stratified target groups) scales.

The ultimate objective is to define the best possible strategy for constructing clinical decision support systems (CDSS) from predictive modelling and analytics. Systems medicine³ approaches offer advantages such as: a) Empowering pervasive computing, integrating inference models, and optimizing data re-use, b) Optimizing the significance of evidences by cross-testing and cross-referencing against datasets, c) Generating patients outcomes by cross-validating large sets of features.

It is necessary the integrability of personal, family, genetic, environmental, social and clinical records, including unstructured information sources, such as imaging, complex treatment plans, and omics experiments. It is desirable the portability of algorithms for linking phenotype and genotype data (see for instance eMERGE, https://emerge.mc.vanderbilt.edu/), expected to benefit from the future growth of biobanks (justifying re-phenotyping) and drug banks (aimed at drug repositioning). EHS-driven genomic research is increasingly founded on clinical relevance⁴ and aimed to re-phenotyping patients. The change of paradigm involves interoperability in EHSs, with a shift toward integrative modeling assessment, sequential decision making, analysis of competing hypotheses, composite testing strategies and other. An important component of EHS is disease ontology (DO).⁵ Essential in many domains, ontologies capture knowledge within a hierarchy of interrelated concepts, and define the basic terms and relations in biological domains. Further expansion of DO in large network- and module/community-driven approaches are now occurring.

Deep Phenotyping

Recently, a variety of factors seen through the lens of human diseases, such as genetic variation, phenotyping,⁶-⁸ drug and epitope data, have led to novel developments such as the integration of ICD-10 (http://www.who.int/classifications/icd/en/), MeSH (https://www.nlm.nih.gov/mesh/MBrowser.html), and OMIM (Online Mendelian Inheritance in Man) resources. The use of the Unified Medical Language System (UMLS) has also taken place from Concept Unique Identifiers for each disease term. Undoubtedly, the exploration of linkage within EHRs may require extending DO to comorbidities, multi-disease maps, epidemiological and social network statistics, implying complex methodological, annotation and validation tasks. EHR phenotyping includes traditional features, such as genetic, clinical, environmental, and others like lifestyle, wellness, healthcare and medical decision making processes.⁹ Lifestyle factors such as diet or physical activity.
and conditions such as overweight, obesity, weight change and smoking are considered modifiable. Their combinations expand the possible definitions of clinical disease subtypes and identifications of disease markers as “deep phenotypes”. In turn, this brings better population stratification and disease characterization. High-throughput phenotyping approaches (phenomics) imply a systematic study aimed at a new disease taxonomy (http://dels.nas.edu/Report/Toward-Precision-Medicine-Building-Knowledge/13284). Leverage goes on data mining, classification and prediction that correlate phenotypes with gene functions, identify signatures and landscapes of disorders, and infer correlative versus causative patterns. The EHS logic is designed exactly for such purposes, increasing the spectrum of common disease features while disambiguating the characteristic disease features.

At least three scenarios are potential marker generators, all dimensions are based on observational factors: a) Those that can be validated in a complete and correct way to justify causal evidence; b) Those needing reproducibility for more substantial validation, thus offering support of only correlative evidence; c) Those limited in replicability and indicating plausibility, justified only by probabilistic assessments. Causality is the key factor for dynamic treatments in chronic diseases involving sequences of decision rules at each intervention point so to allow for therapy personalization. Confounding factors are likely present in EHS, leading to spurious association of evidences of any nature, due to genetic, environmental, demographic factors. For instance, genetic loci associated with diabetes risk influence insulin secretion and likely also neoplastic behavior, but the potentially confounding role of nutrients requires consideration.

**Multi-Evidence Precision Medicine**

Big Health Data present inherently structural complexity, one which is difficult to tackle. Phenomena like entanglement in quantum or equilibrium statistical physics, visualization in computational sciences, deconvolution and compressive sensing in information theory, can respectively describe, represent or decipher complexity to a certain extent. When data come into play, often the goal is to find some sufficient statistics, thus retaining full data information from only a reduced set of it (without loss). But what does it really mean in EHS, what’s the optimal reduction from a high-dimensional and partially undefined information space? And, how can we assess sufficiency in a context of technological limitations, highly variable experimental power, objective and subjective evidence types, inhomogeneous scale/g regularity of measurements? Consequently, how measurable are necessary statistics, those computable from any sufficient statistic without reference to the original data? Contextual information is key to compensate for information paucity. For instance, gene expression analysis remains a great source of insight into biological variation but both its influence on regulation cannot be confined to signal-to-noise ratios. Whether non-coding RNAs are pervasively transcribed or not along the genome that Next Generation Sequencing methods allow to deeply screen, depends on how precisely we assign significance to the detected biotypes, and functionally validate them, something having only limited possibilities of solution without the recourse to multiplexed approaches.

A double paradox is at play: precision medicine would require specificity for components, say drug targets, disease markers, etc., for which not enough information might be available unless each component is referred to a system, which then may become too complex due to its dynamics. The classical bias-variance tradeoff becomes very relevant as the approximation algorithms aimed to learn in big data systems are expected to include errors due to uncertainty from system’s variability and other sources. These additional errors should be minimized to maximize precision, leading to the usual decomposition into squared bias and variance. The simultaneous reduction of both components implies a tradeoff between model complexity (the best possible) and model selection (the most feasible). Learning in EHSs is therefore a process dependent on how precisely the algorithms adapt to the complex data, i.e. with low bias, but possibly high variance. This is the case with heterogeneous training data typical of problems such as molecular profiling of cancer subtypes or detection of cell sub-populations resistant to therapies. Also, several new anticancer agents are currently targeted to driver ensembles, rather than individual ones, and this through integrative approaches at a system’s scale. All factors involved in inter-cellular and inter-organ communication, together with their expected changes, will carry crucial information regarding the insurgence of diseases. The ‘cellular communicome’ encompasses the proteins within an organism which carry information from one cell to another, such as cytokines, hormone-like proteins, growth factors and so forth. An advanced conceptualization of the disease process as a rewired communicome makes easier the understanding of how the shared repertoire or the bowtie signaling structure creates comorbidities (Figure 2). These latter are observed through patterns at population scale, but their trajectories are individual ones when it comes to establish onset of the disease and age at the time of onset, which calls for the need of incorporating probabilities of early events in development, causal to the insurgence of diseases or. Since causal inference plays a key role in drawing personalized comorbidity maps including information on development, such as models on fitness landscape dynamics during development, theoretical and applied statistical work are needed to hamper the precision. ComoR is an R package that computes estimators of the disease comorbidity associations, starting
from initial diagnosis, employing genetic and clinical data of a patient, and delivering the risk map of disease comorbidity. It makes use of different causal inference packages, including CytoCom, an interactive Cytoscape plugin to search, explore, analyze and visualize human disease comorbidity networks.

**Marker Re-Modulation**

Complex diseases involve multifactorial markers (due to comorbidities, therapies, etc.). For instance, rheumatoid arthritis (RA) is an inflammatory autoimmune disease of partially unknown etiopathogenesis, directly (inherent risk factor) and indirectly (associated with atherosclerosis) increasing the risk of death for cardiovascular disease. Chronic inflammation plays a major role by negatively acting on the artery and inducing the re-modulation of other risk factors (smoking, diabetes mellitus, obesity, hypertension, insulin resistance, metabolic syndrome), and is naturally associated with changes in gene expression levels, suggesting possible links with RA heterogeneity and progression stages which remain still not clearly distinctive in their diagnostic, prognostic and predictive power with respect to therapy. High-genomic content EHs could reveal extremely useful to cover current knowledge gaps, even if in some contexts consistency is hard to achieve, like with RA-insulin resistance conditions, or with LDL cholesterol levels considered before and after statin therapy, for which it remains unclear the specification of a role in establishing cardiovascular risk reduction. In type 2 diabetes (T2D) major roles are played by both environmental and genetic factors, with genetic loci still contributing only to a certain extent to heritability. GWAS identified less than a hundred SNPs associated to T2D, and EWAS only five. Algorithms have been proposed to identify with accuracy T2D cases and controls within GWAS. Epidemiological and clinical studies point to gene-environment interactions or epigenetic changes regulating gene expression levels for explaining T2D. Among the known markers, subclinical inflammatory events (secretion of TNF and pro-inflammatory cytokines) represent early signals, while other regulatory mechanisms are associated to transcription factor NFkB, circulating microRNAs, and epigenetic modifications affecting gene expression (histone hyperacetylation and chromatin remodeling), clearly important for treatment purposes. Novel biomarker discovery will also benefit from comprehensive screenings (see the Genetic Association Database survey of ORFs of the human dark matter proteome for evidencing associations with neoplasms leading to the discovery of new targets for diabetes, and also in diabetes-associated disorders, thus suggesting novel therapeutics.

**Discussion**

Novel uses of social media are inspiring completely new assessments of health and disease, in terms of interactions (care and cure), perceptions of conditions (self-quantified), remote delivery or monitoring (telemedicine), and community models. Collectively, these will open up a wealth of opportunities.

**FIGURE 2: Comorbidity DSS.** A general frame for hierarchical classification of risk into classes (low-high) and communities, adding further granularity in the patient stratification by making each risk class subjected to assessment and calibration. Symbols: T2D: type 2 diabetes mellitus; CAD: coronary artery disease; RA: rheumatoid arthritis; COPD: chronic obstructive pulmonary disease.

**BOX 2. Disease reversibility.**

A few years ago, the world news reported about the UK Doctor Starts saying: ‘I’d Rather Have HIV Than Diabetes’ (http://www.spectator.co.uk/features/9185591/why-id-rather-have-hiv-than-diabetes/). Now, T2D reversibility is changing the scenario. Recent studies have reported that the risk for diabetes may originate from metabolic stress during intrauterine and early postnatal development. This may involve subtle epigenetic changes. The current point of view is that T2D is not inevitably progressive and life-long; if the insulin secretion is not lost completely, it can be reverted. Reversibility may be due to a large spectrum of factors including nutrition, exercise, drugs, infections etc. One challenge is the identification of markers of disease trajectories, signals of evolving entities, aimed to personalized programs that reverse pre-diabetes and diabetes conditions. RA and T2D are independent risk factors for ischaemic heart disease, but also convoluted diseases (Figure 3). Both T2D and RA cross-talk with bone-related pathologies, particularly with infections and aging. RA appears to be associated with an increased risk for T2D in some populations. Conversely, recent studies have highlighted an elevated risk of RA in female Taiwanese patients with T2D.
Emerging interest focuses on nutrition, often found to be the smoking gun for disease association. The field of nutritional epidemiology could benefit from EHS-driven investigations, after the announcement of its death (see Willett’s presentation “Webcast of diet and cancer: status report in 2014”, at the 2014 AACR annual meeting). The estimate first circulated in 1981 of about one third of cancers being due to nutrition is likely correct, but nutrition would be less important than obesity and inactivity, although nutrition is clearly involved in both. Nutrition’s role can be better elucidated and cleared from biases, and with diet it is a key factor to enable modifications in epigenetic risk markers.

**Acknowledgments**

We would like to thank colleagues from our Institutes, and active in our own fields or other research environments, who have shared for years with us ideas and analyses on translational medicine, experimental and clinical investigation aspects, model strategies and high performance computing problems.

**Box 3. Key Messages:**

- Electronic Health Systems represent complex junctions of phenotypes
- Repurposing phenotypes will be central to Precision Medicine
- Are ‘sufficient statistics’ computable from Electronic Health Systems?
- Are ‘necessary statistics’ useful to Precision Medicine?

**Enrico Capobianco** (PhD, Statistics) is Lead Scientist at the Center for Computational Science of the University of Miami, and associate with the Institute of Clinical Physiology of the National Research Council in Italy. He has a vast international research experience in Cancer Genomics, Systems Medicine, Network Science, and Big Data.

**Pietro Lio’** is Reader in Computational Biology at the Computer Laboratory, University of Cambridge. He has obtained a Doctorate in Theoretical Genetics from the University of Pavia, a Doctorate in Non Linear Dynamics and Complex Systems from the University of Firenze and a Honorary Master from the University of Cambridge. In October 2013 has received the Lagrange Fellowship (Lagrange Foundation, ISI, Turin). He has more than 250 peer reviewed publications.
References


For more information: personalizedmedicinecoalition.org
202.589.1770
1710 Rhode Island Ave, NW, Suite 700
Washington, DC 20036

The Personalized Medicine Coalition, representing innovators, scientists, patients, providers and payers, promotes the understanding and adoption of personalized medicine concepts, services and products to benefit patients and the health system.