Review

Why we need artificial intelligence in environmental rheumatology

E. Grossi

Villa Santa Maria Foundation, Tavernerio, Italy.

Enzo Grossi, MD

Please address correspondence to:
Enzo Grossi
Scientific Director,
Villa Santa Maria Foundation,
Via IV Novembre,
22038 Tavernerio, Italy.
E-mail: enzo.grossi@bracco.com

Received on May 24, 2024; accepted in revised form on May 27, 2024.

J Environ Rheumatol 2024; 1: 10-14.

Key words: Artificial intelligence, environment, rheumatology

Competing interests: none declared.

ABSTRACT

Artificial intelligence (AI) holds the potential to address various challenges in environmental rheumatology. These challenges include the need for better understanding complex relationships within environmental data, fast prediction of environment-related parameters, risk assessment, and environmental decision-making. AI can assist in analysing large datasets to identify genetic risk factors for adverse reactions to toxic exposures, as well as in detecting gene-environment interactions and epigenetic alterations. Furthermore, AI can play a crucial role in toxicity prediction by utilising deep learning algorithms to assess complicated bioactivity data, offering a solution where traditional methods fall short. Despite the promise of AI, challenges such as insufficient or poor-quality data, model interpretation difficulties, and concerns about bias in outcomes need to be addressed for successful integration of AI in environmental rheumatology.

Introduction

The recent launch of the *Journal of Environmental Rheumatology* offers the cue to deepen the hot topic of the role of artificial intelligence in environmental medicine regarding rheumatology. Adding environment information to classical clinical information means two things: an increase in quantity and in variety of data, two fundamental features of what we call big data (1).

The new challenge for epidemiology in our century is to deal with the tsunami of data, the so-called big data, covering the spectrum of genomic, molecular, clinical, epidemiological, environmental and digital information. The fusion of data from all these sources has in itself all the potential to influence decision-making processes of the individual physician and, more generally, in public health (2).

The first phenomenon that advanced the frontier of Big Data in health is the development of biotechnology employed in the field of so-called 'omics' sciences, disciplines that have as their object a very in-depth study of the cell, through a detailed analysis of biological processes observed at different levels going from genes (genomics), their functions (functional genomics), their DNA transcripts, *i.e.* RNA (transcriptomics), the proteins (proteomics) and the metabolites that found in the body (metabolomics), the ways in which molecules interact (interactomics), interactions with the gut flora (microbiomics) and how they can modify the DNA (epigenomics).

The omics family is enlarging with the chronobiomics. In recent years, a new construct has emerged, that of the so-called chronobiome, term first coined in 2017 (3), based on a deep multi-omics characterisation of an individual that outlines an individual pattern of molecular networks influenced by behaviour, environmental stimuli, or stressors. Since the molecular circadian clock coordinates the body's rhythms for each individual person, symptoms of disease, the effectiveness of treatment and many physiological processes in the body vary depending on the time of day. The evidence that circadian misalignment due to artificial light, shift work, and jet travel is common in modern life and contributes to a wide range of human diseases is strong (4). Despite our understanding of circadian rhythms is expanding to provide molecular insights into physiology and

disease, yet the challenge remains to translate these insights into the practice of clinical medicine integrating with other omics disciplines.

These sophisticated technologies of molecular biology technologies lead to an extremely high volume of data of about 50 terabytes in a single subject (5).

The second phenomenon is the emergence of exposomics, more directly related to the environment, a new, strongly multidisciplinary discipline, which began making headlines and promises to be a fundamental approach to understanding age-related chronic degenerative diseases, like most rheumatic diseases are.

A tangible definition of the exposome was proposed by Miller and Jones in 2014: "The cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment in term of climate and pollutants, from diet, from social behaviour, and from endogenous processes" (6).

Unlike the genome, exposures are transient and vary on both short- and long-term time scales, making quantitative assessment challenging.

The first historical trace of the scientific interest in environmental related diseases can even be found in the writings of Bernardino Ramazzini. His treatise '*De morbis artificum diatriba*' published in Padua in 1708, the first study in the history of medicine on occupational diseases, is considered the founding act of what is now called occupational medicine. Ramazzini examined and analysed the context of working conditions and the diseases resulting from them, of many occupations (40-50) and, in addition, described the possible health risks related to each job and their possible remedies (7).

In addition to this, he considered, supplementing the data already obtained, the climatic conditions in which these jobs were or could be carried out. The relationship between risks and disease observed anticipated the scientific method, still used today, based on epidemiological studies (8).

It was only due to the availability of reliable analytical techniques that exposure science was able to develop in the 1950s and 1970s as a complement to occupational medicine by investigating the external environmental concentration of chemicals that can enter the body through inhalation, ingestion, or skin contact.

In the 1990s, internal biomarkers of exposure to toxicants began to be measured, and only later the levels of incriminated chemical compounds were measured. The real paradigm shift took place in 2005 when Christopher Wild first introduced the term 'Exposome', as opposed to 'genome' to mean everything a human being has been exposed to throughout his or her life, since conception through the external environment, and internal environment (9). This in the knowledge that since the description of the human genome, the results of the very expensive genome-wide association studies had failed to explain most of the variability in human diseases.

Why is it important to assess exposure? As the Californian scientist Stephen Rappaport explained in a famous article published in the journal *Science* in 2010, if one adopts a rather broad definition of the environment, a large part of the increased risk of cancer and degenerative diseases is mainly related to environmental rather than genetic factors (10). It is therefore a very powerful concept, capable of explaining more than 90 per cent of the likelihood of encountering chronic degenerative diseases typical of old age.

The big problem is that exposures are highly dynamic and therefore difficult to master. Concentrations of exogenous or endogenous chemicals vary over time between people and populations. The scale of variability is enormous and so the question arises as to how we can be sure that we are focusing on truly important chemicals or periods.

Incorporating exposure biomarkers into population studies

To take advantage of biomarkers, we need to think about the design of population studies by looking at diseases, and not markers, as endpoints. Every study design has something to offer, but we need to think carefully about the questions to ask in a particular type of design. For example, cross-sectional studies (studies of groups with different characteristics at a given time) usually allow a focus on a few people with a great deal of detail, collect a lot of exposure data, and determine what additional information is needed to validate relationships between hypothesised exposure biomarkers. Case-control studies can look at exposures that have occurred recently if biological samples are inadequate. Cohort studies (studies of a group with a common set of characteristics over time) are notoriously the 'crown jewels' in the armamentarium of epidemiology, but financial constraints often limit the collection of multiple biological samples.

The different study designs complement each other, and scientists should consider how to integrate them or use them in tandem to get a better picture of exposure. For example, Rothman and his colleagues at the National Cancer Institute and the University of California, Berkeley, used a series of cross-sectional studies to assess biomarkers of benzene exposure in workers. The studies helped develop hypotheses that were later tested in cohort studies designed to follow workers over the course of the disease. Rothman also suggested applying the same analytical tools in studies of different types or classes of exposure (11).

Most biomarkers require large amounts of biological material, such as blood and urine, and can therefore be difficult to use in cohort studies.

As an alternative, a mixed study design can be used in which the expensive biomarkers to be evaluated are measured in only a subset of samples and less expensive measurements are made on all samples. The more expensive instrument is then used to calibrate the less expensive one.

It is desirable that in the future, different scientific fields such as social, nutritional, and environmental science use similar techniques and technologies by building an interconnection between the different fields and encouraging the sharing of biomarkers, questionnaires, and other research tools.

Of course, this growth in technology

inevitably creates a huge computing challenge. The real problem for exposomics is having to manage such a large number of different kinds of data, without easily having adequate tools to process them. Managing information has now become a heavy burden for those involved in public health.

Table I. Motivations to apply complex mathematical systems in environmental rheumatology.

- Health processes are based on networks of endogenous and exogenous elements interacting in a complex way.
- Health status is the consequence of dynamic processes that regulate these networks.
- Nonlinear critical thresholds bring to pathology.
- The predictions must be applied at individual patient level.
- Huge amount of data per subject hamper classic statistical tests.

Living system require the mathematics inherent to AI

A breakthrough of the twentieth century, which has been facilitated by computer science, has been the recognition that simple rules not always lead to stable order, but in many circumstances instead lead to an apparent disorder characterized by marked instability and unpredictable variation for reasons intrinsic to the rules themselves. The phenomenon of rules causing emerging disorder, counterintuitive to many people, is well described in chaos theory, also called nonlinear systems theory, which provides new insights into processes previously thought to be unpredictable and random (12, 13). It also provides a new set of tools that can be used to analyse physiological and clinical data. The advancement of knowledge and the progress of understanding the nature of bodily rhythms and processes have shown that complexity and nonlinearity are ubiquitous in living organisms. These rhythms arise from stochastic (involving or containing a random variable or variables), nonlinear biological mechanisms interacting with fluctuating environments.

There are many unanswered questions about the dynamics of these rhythmic processes: for example, how do the rhythms interact with each other and the external environment? Can researchers decode the fluctuations in physiological rhythms to better diagnose human disease? Mathematical and physical techniques combined with physiological and medical studies are addressing these questions and are transforming our understanding of the rhythms of life.

Other researchers as McEwen and Wingfield have introduced the concept of allostasis, *i.e.*, maintaining stability through change, as a fundamental process through which organisms actively adjust to both predictable and unpredictable events. Allostatic load refers to the cumulative cost to the body of allostasis, with allostatic overload being a state in which serious pathophysiology can occur. In this regard chaos theory seems to fit quite well with biological adaptation mechanisms (14).

The emergence of complexity in rheumatology

Non-linearity, complexity, fuzzy interaction are emerging features of chronic rheumatic diseases which account for most morbidity in western world. Unfortunately, even the most powerful and well-established statistical methods were developed in the first half of the past century when the scenario was dominated by acute infective diseases and the available information was much simpler, or at maximum "complicated" rather than "complex" comparison with today.

More features and more variety in data imply more information and potentially higher accuracy. Unfortunately, more features we have, the more difficult information extraction is. In this high dimensional space, the hyper points corresponding to single individuals are sparse and the notion of proximity fails to retain its meaningfulness. For this reason, clustering become extremely hard to be performed. In this situation we are dealing with flat, rectangular data set, a sort of telescope data set. This kind of data set are intractable from a traditional statistics point of view since the excessive number of degrees of freedom allows any kind of data interpolation most of the time meaningless.

Complexity is an adaptive process; it is time sensitive, and over time complex processes evolve and/or degenerate. Complexity is based on small elementary units working together in small populations of synchronous processes.

In a complex system each component changes, over time, losing its identity outside of the system. Complexity needs a different kind of mathematics, able to handle chaotic behaviour, nonlinear dynamics, and fractal geometry (15).

Computational and mathematical medicine needs different statistical approaches, based on new mathematical and logic assumptions broadly belonging to complex theory setting allow to tame these intractable data sets. Seen in this perspective computer science and artificial intelligence are now playing the role which mathematics did from the seventeenth through the twentieth centuries: providing an orderly, formal framework and exploratory apparatus for knowledge progress. Actually, the coupling of computer science and these new theoretical bases coming from complex systems mathematics allows the creation of "intelligent" agents able to adapt themselves dynamically to problem of high complexity: the machine learning systems (ML).

Machine learning systems can adapt themselves to a dynamic environment, and time for them is not "noise" but is rather a way to reduce potential errors. There are several different reasons to apply complex systems mathematics on predictive medicine and some of them are listed in the Table I.

The contribution of artificial intelligence in environmental rheumatology

The last forty years have seen a revolution in the field of artificial intelligence to support clinical medicine. The 1960s, 1970s and 1980s were characterised by the prevalence of so-called expert systems, based on the applica-

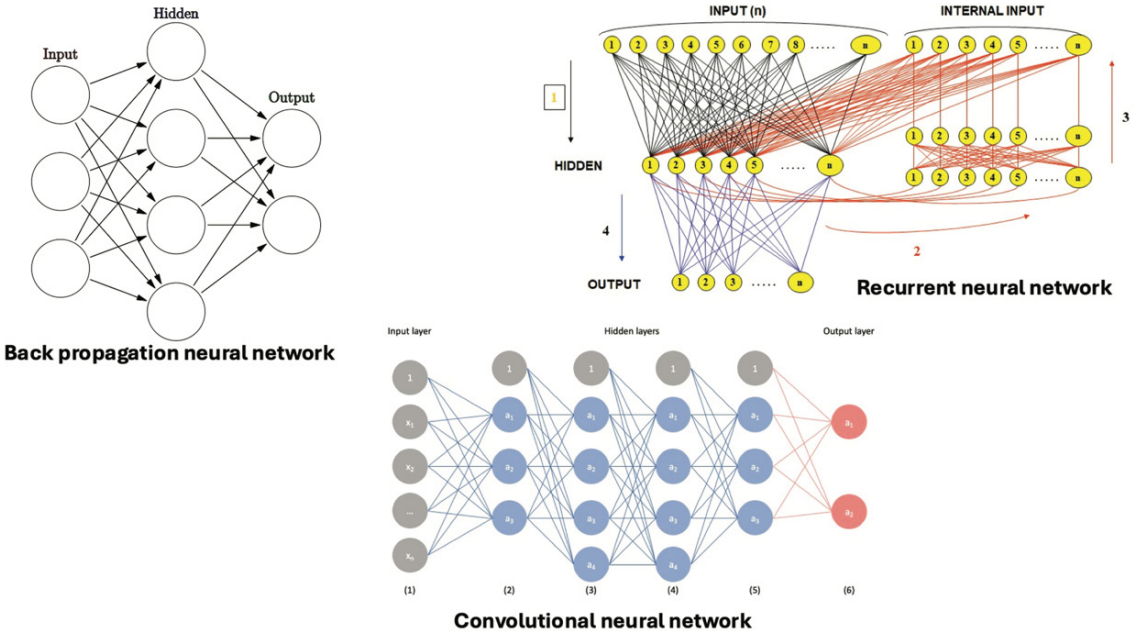

Fig. 1. Architecture of artificial neural networks milestones.

tion of if-then-else instructions and the acquisition of knowledge from subject-matter experts. But after a period of euphoria and optimism about the potential of such systems, culminating in the 1980s, it became clear that they manifested important fragilities arising from their difficulty in taking into account the complex, and often implicit, contextual conditions that come into contextual conditions that come into play in the medical field, and that physicians generally deal with by using common sense, which is precisely most difficult to achieve through inductive/deductive logic.

The phase of evolution following to the probabilistic led to artificial neural networks having as milestones the back-propagation technique from Rumelhart in 1987 (16), the recurrent neural networks (17), which constituted a decisive advance in sequence modelling, such as those of texts, and eventually convolutional networks (18) structured in 'cascades' of increasing of generalization, according to a scheme borrowed from animal visual neuronal systems, and for this reason characterized by very fine pattern recognition potentialities of which the architecture is rich. These neural networks led to the definition of so-called 'deep learning' from Google (19). Figure 1 shows the architecture of the systems listed in the previous paragraph.

Artificial intelligence (AI) is increasingly being integrated into rheumatology to enhance diagnostics, disease monitoring, and treatment personalization (20, 21). AI techniques, including machine learning and deep learning, are revolutionizing the field by aiding in disease classification, prediction, prognosis, and treatment response analysis for autoimmune rheumatic diseases like rheumatoid arthritis and systemic lupus erythematosus. Deep learning, in particular, shows promise in interpreting unstructured data such as images and text, enabling the detection of joint erosions, predicting disease activity, and identifying specific features like halo sign on ultrasound in rheumatology. Moreover, healthcare professionals are increasingly optimistic about AI's potential benefits in managing patients with rheumatic diseases, with a focus on improving patient management through accurate prognosis and treatment personalisation. It is quite clear that AI and ML can contribute to the understanding of how environmental exposures over a lifetime (known as the exposome) drive the inception and the severity of rheumatic diseases and how we can use this knowledge for primary and secondary prevention.

The advantage of AI is that it can enhance its learning base through continuous learning, which allows a long-trained AI device to react and make decisions in a short time. At the same time, the programmed operation of AI allows it to adjust its cognition of different environments in a self-iterative way (22). Therefore, these characteristics of AI make it particularly outstanding in the aspect of continuous monitoring. Compared with traditional methods that rely on manual sampling in the field, AI can do real-time monitoring for a long time. This makes it more informative for researchers. For example, researchers can summarise the change trend of detected substances in the environment based on artificial intelligence data, which enables researchers to adjust research strategies in a timely manner. In terms of pollution treatment, researchers can know the main sources of pollution and take timely measures based on the discharge data of surrounding pollutants (23).

Where we are

To answer to this question, a search on the keywords environment & rheumatic diseases & artificial intelligence was undertaken on MEDLINE (via PubMed), The Cochrane Library, Google Scholar, from databases inceptions until April 2024. The number of papers retrieved with the keywords "environment & rheumatic diseases" without inclusion criteria was 831. The number of papers satisfying as inclusion criteria "clinical studies", excluding microbiota, was 4 (Table II).

Table II. Published study on environmental factors and rheumatic diseases.

Author	Year	Reference	Journal	Type of study	No. of subjects	Type of environmental factors
Hawley & Wolfe	1994	22	*Pain*	Cohort study	2,523	Light and season
Mori *et al*.	2019	23	*BMC Musculoskelet Disord*	Cohort study	12,839	Season
Nilssen *et al*.	2020	24	*J Rehabil Med*	RCT	64	Warm climate
Herly *et al*.	2020	25	*Sci Rep*	Cohort study	160	Season and vitamin D

No study was retrieved by adding "artificial intelligence" to previous key words.

From this review it seems that there is a huge space of improvement in our present knowledge about the classical environmental factors and rheumatic diseases.

In conclusion, since as mentioned genetic variations probably play a relatively minor overall role in rheumatic diseases, we should devote more attention to the endogenous and exogenous environmental effectors driving the epigenetic control with the subsequent genes switch on-off in contrast to the limited value of genetic by herself without such control. However, since examples of application of AI to environmental rheumatology are still scarce waiting for the development of this new discipline we need to move from environmental monitoring to bio-monitoring and use biomarkers to identify and process the exposome. Developing exposomics will require extraordinary efforts in many disciplines. It will require inputs from environmental toxicology, epidemiology, molecular biology, epigenetics, analytical chemistry, chemometrics, bioinformatics and engineering, and finally, mathematics, still largely disconnected disciplines for which a common language will need to be developed. This language will come from AI.

References

1. MOONEY SJ, WESTREICH DJ, EL-SAYED AM: Epidemiology in the era of big data. *Epidemiology* 2015; 26(3): 390-4. https://doi.org/10.1097/ede.0000000000000274
2. KHOURY MJ, LAM TK, IOANNIDIS JP *et al*.: Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiol Biomarkers Prev* 2013; 22(4): 508-16. https://doi.org/10.1158/1055-9965.epi-13-0146
3. SKARKE C, LAHENS N, RHOADES S *et al*.: A Pilot Characterization of the Human Chronobiome. *Sci Rep* 2017; 7(1): 17141. https://doi.org/10.1038/s41598-017-17362-6
4. ALLADA R, BASS J: Circadian Mechanisms in Medicine. *N Engl J Med* 2021; 384(6): 550-61. https://doi.org/10.1056/NEJMra1802337
5. TOPOL EJ: Individualized Medicine from Prewomb to Tomb. *Cell* 2014; 157(1): 241-53. https://doi.org/10.1016/j.cell.2014.02.012
6. MILLER GW, JONES DP: The nature of nurture: refining the definition of the exposome. *Toxicol Sci* 2014; 137(1): 1-2. https://doi.org/10.1093/toxsci/kft251
7. RAMAZZINI B: The art of medicine is similar to navigation. 5th lecture. Held on 4 November 1703. *Med Lav* 2010; 101 (Suppl 3): 54-65.
8. FRANCO G: Ramazzini and workers' health. *Lancet* 1999; 354(9181): 858-61.
9. WILD CP: Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 2005; 14(8): 1847-50. https://doi.org/10.1158/1055-9965.epi-05-0456
10. RAPPAPORT SM, SMITH MT: Environment and disease risks. *Science* 2010; 330(6003): 460-1.
11. ROTHMAN N, SMITH MT, HAYES RB *et al*.: An epidemiologic study of early biologic effects of benzene in Chinese workers. *Environ Health Perspect* 1996; 104 (Suppl 6): 1365-70. https://doi.org/10.1289/ehp.961041365
12. FIRTH WJ: Chaos, predicting the unpredictable. *BMJ* 1991; 303(6817): 1565-8. https://doi.org/10.1136/bmj.303.6817.1565
13. GROSSI E: Chaos theory. *In* KATTAN MW, COWEN ME (Eds.): Encyclopedia of medical decision making. Sage Publications, 2010: 129-132.
14. MCEWEN BS, WINGFIELD JC: The concept of allostasis in biology and biomedicine. HORM BEHAV 2003; 43(1): 2-15. https://doi.org/10.1016/s0018-506x(02)00024-7
15. GROSSI E: Medicine and Mathematics of Complex Systems: An Emerging Revolution. *In* CAPECCHI V, BUSCEMA M, CONTUCCI P, D'AMORE B (Eds.): Applications of Mathematics in Models, Artificial Neural Networks and Arts. Springer, Dordrecht 2010: 415-30. https://doi.org/10.1007/978-90-481-8581-8_15
16. RUMELHART DE, HINTON GE, WILLIAMS RJ: Learning internal representations by error propagation. *In*: RUMELHART DE, MCCLELLAND JL (Eds.): Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1, MIT Press, Cambridge, Massachusetts, 1987: 318-62.
17. HOCHREITER S, SCHMIDHUBER J: Long Short-Term Memory. *Neural Comput* 1997; 9(8): 1735-80. https://doi.org/10.1162/neco.1997.9.8.1735
18. LECUN Y, BOSER B, DENKER JS *et al*.: Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput* 1989; 1(4): 541-551. https://doi.org/10.1162/neco.1989.1.4.541
19. LECUN Y, BENGIO Y, HINTON G: Deep learning. *Nature* 2015; 521(7553): 436-444. https://doi.org/10.1038/nature14539
20. QUIDWAI S, KIRBY C, MURPHY G: AI Applications in Rheumatology. *In* BYRNE MF, PARSA N, GREENHILL AT, CHAHAL D, AHMAD O, BAGCI U (Eds.): AI in Clinical Medicine: A Practical Guide for Healthcare Professionals. Wiley Press 2023. https://doi.org/10.1002/9781119790686.ch22
21. MCMASTER C, BIRD A, LIEW DFL *et al*.: Artificial Intelligence and Deep Learning for Rheumatologists. *Arthritis Rheumatol* 2022; 74(12): 1893-905. https://doi.org/10.1002/art.42296
22. GROSSI E: Artificial Adaptive Systems and predictive medicine: a revolutionary paradigm shift. *Immun Ageing* 2010; 7 (Suppl. 1): S3. https://doi.org/ 10.1186/1742-4933-7-S1-S3
23. MANISALIDIS I, STAVROPOULOU E, STAVROPOULOS A, BEZIRTZOGLOU E: Environmental and Health Impacts of Air Pollution: A Review. *Front Public Health* 2020; 8: 14. https://doi.org/10.3389/fpubh.2020.00014
24. HAWLEY DJ, WOLFE F: Effect of light and season on pain and depression in subjects with rheumatic disorders. *Pain* 1994; 59(2): 227-234. https://doi.org/10.1016/0304-3959(94)90075-2
25. MORI H, SAWADA T, NISHIYAMA S *et al*.: Influence of seasonal changes on disease activity and distribution of affected joints in rheumatoid arthritis. *BMC Musculoskelet Disord* 2019; 20(1): 30. https://doi.org/10.1186/s12891-019-2418-2
26. NILSSEN IR, KOKSVIK HS, GRØNNING K, STEINSBEKK A: Rehabilitation in warm climate for young adults with inflammatory arthritis: A 12-month randomized controlled trial. *J Rehabil Med* 2020; 52(4): jrm00040. https://doi.org/10.2340/16501977-2666
27. HERLY M, STENGAARD-PEDERSEN K, VESTERGAARD P *et al*.: Impact of season on the association between vitamin D levels at diagnosis and one-year remission in early Rheumatoid Arthritis. *Sci Rep* 2020; 10(1): 7371. https://doi.org/10.1038/s41598-020-64284-x