ECRONICON OPEN ACCESS

EC PHARMACOLOGY AND TOXICOLOGY
Review Article

Health and Medical Analytics, through Analytical Focus and Contextualization, with New Challenges and Opportunities in the Context of Big Data

Fionn Murtagh*

Director, Centre of Mathematics and Data Science, University of Huddersfield, Huddersfield, UK

***Corresponding Author:** Fionn Murtagh, Director, Centre of Mathematics and Data Science, University of Huddersfield, Huddersfield, UK.

Received: May 21, 2019; **Published:** May 31, 2019

Abstract

The central methodology here is Geometric Data Analysis, an alternative term for Correspondence Analysis, for analytics of processes and behaviours. The geometry of the factor space expresses semantics and implicit, underlying, relationships between what is observed or recorded and the attributes or variables that characterize them, and to this there may very well be clustering and that can very often be hierarchical clustering. This implies the processing here of data and information for analytics of processes and behaviours. Here, such processes and behaviours can be patient diagnosis and treatment, and also all that relates to health and that can give rise to the need for medical treatment. Also employed are practical applications from the work of eminent social scientist, Pierre Bourdieu. Also, at issue is the addressing of new societal challenges, and new themes and topics, problems and challenges, in medicine and in health and life sciences. Often achieved is to have relatively acceptable computational complexity.

Keywords: *Qualitative and Quantitative Data Analysis; Data Mining; Unsupervised Classification - Cluster Analysis; Exploratory Data Analysis*

Introduction

This paper describes the focus of analysis, contextualizing the analysis and interpretation, visualization in biplots, and ranking important data components.

In the first chapter of Zhang, *et al.* [1] entitled "Big data and clinical research: perspective from a clinician", author Zhongheng Zhang counterposes, as research, interventional analysis, which is, in fact, experimental research, relative to observational studies. For the former, typically at issue are randomized controlled trials. Also, for the former, selection is required but then there might be bias associated with what is done. An important point made is that patients' treatments are usually complicated by the patients' comorbidities. So, it becomes so very important to have and to use ancillary and contextual observations also. This amounts to the practical setting for big data clinical trials. Such big data, here including in particular, the medical records, may very well include also, demographic attributes, microbiology information on the patient, and other data sources. So, it is noted that such observational data, encompassing what amounts to big data clinical trials, can be very relevant to the "real-world", i.e. detailed and comprehensive information on the patient.

A repository, entitled Medical Information Mart for Intensive Care III, MIMIC- III, with its data on over 40,000 patients is described. Access to that is described. An interesting statement is that the demographics of China will lead to very high-quality big data sources in China. A figure illustrates how a hospital bed is set up for big data recording. In this second chapter entitled "Big data and clinical research: focusing on the area of critical care medicine in mainland China", this figure has in its caption, "In Jinhua Municipal Central Hospital, every bed is equipped with a monitor... and a computer... so that data on vital signs (e.g. respiratory rate, heart rate, blood pressure and

body temperature) and other physiological signals (e.g. extravascular lung water when transpulmonary thermodilution measurement is performed) can be automatically... stored in the electronic medical record (EMR) system...". Description is provided (in the fifth chapter, entitled "Release of the National Healthcare Big Data in China: A Historic Leap in Clinical Research") of an important release at the beginning of 2017 of the "National scientific data sharing platform for population and health (NSDSPPH)" in China, comprising observed and recorded data with 280 million observations or records. This is noted as a very major advance in clinical research.

In the paper here, Section 2 is for how Big Data can be availed of, to form the context for the patient's treatment. There is touching on the bias that can come from self-selection of behavioural and other lifestyle data; proposing is the importance of "bridging" and shared patterns and associations in the data. Hence, this is to benefit from the methodology of eminent social scientist, Pierre Bourdieu.

Section 3 is the orientation of the analytical work undertaken, and having contextual data and information, for helping interpretation following the analyses and in particular, the causation or underpinning that explains the principal analyzed data and linked information. There is here an early stage of mental health analytics.

Big data methodology developments

This section describes the importance for Big Data analytics, and associated paradigm shifts. Then there is the current theme of Open Data and lots of other sources, of data and associated or related information. The analysis here uses the Correspondence Analysis factor space mapping and hierarchical clustering expresses and is a mapping of the complex context of all that is at issue, cf. Murtagh [2]. It can be noted how Data Science requires the integration of data and of analysis, and furthermore it is interdisciplinary and multidisciplinary.

Keiding and Louis [3] is an important perspective on how Big Data analytics can require changes in the statistical foundations relating to the population distributions and also how sampling functions. One very central theme is: "There is the potential for big data to evaluate or calibrate survey findings". Under discussion is, "how data... tracks well with the official", i.e. standard practice and standard procedures.

Stated and discussed is how there will be "the value of using 'big data' to conduct research on surveys (as distinct from survey research)". This implies that survey data is to be calibrated, using Big Data. Limitations though are clear: this means bias, and in Keiding and Louis [3], this results from self- selection of survey respondents, for example, in surveys in social media, and possibly also this can be relevant for health and lifestyle data sources.

To address the need for Big Data calibration here, this statement was used: "When informing policy, inference to identified reference populations is key". So, in effect, there is the "bridge" or link needed, between the analytical processing carried out and the way in which there will be calibration, hence relating observed data sources to particular and highly relevant context and basis. Furthermore, this must be linked also to decision-support.

An important aspect of what is at issue here is data aggregation, that can be termed also, because of how data aggregation is carried out, averaging or representation, etc. Such data aggregation may be then viewed as replacing individual and even personal attributes, details, etc. and having focus of attention on what is representing all. In Keiding and Louis [3], this is to be addressed, first, through the relating of given data sources with Big Data sources, and based on that, to have such linkage giving rise to interpretable, explanatory, underpinnings. This expression, the need for the "formulation of abstract laws" that bridge sampled data and calibrating Big Data, this is carried out for the data analyst and for the application and domain specialist, as mathematically based, as geometric and topological. The latter term implies hierarchical properties.

Pierre Bourdieu's social science work involved such concepts as field and homology. As described in Lebaron [4], Bourdieu's work implied global effects, that can be understood here as being related to Big Data sources. For the latter, there is its conceptual structuring which is a great deal beyond the data source that one is analyzing. Bourdieu's terminology included homology, which can be described

as repeated and shared patterns, and another term in Bourdieu's work is field. That latter can be described as the factor plane, or more than one plane in the factor space that is the mapping of the qualitative and quantitative data carried out by Correspondence Analysis. Displaying the factor plane, or various planes, is a visualization treatment in the analysis carried out. From the mathematical point of view, the mapping into the Euclidean metric endowed factor space is well expressed also as Geometric Data Analysis, i.e. the geometry and its properties of the data sources.

Analytical orientations and their contextualization

Our interest in the data source here is in regard to mental health. Due to the analysis first stage and second stage, changes were made in what constitutes the main analytical data, here variables, and then what constitutes the supplementary variables. The latter are mapped into the factor space, when established. The supplementary elements, i.e. variables or recordings, that are, respectively, columns or rows in the data array/matrix that is being examined, i.e. analyzed; such supplementary columns constitute the context information.

HSCIC [5] provides an adult psychiatric morbidity survey, in England, in 2007. In this questionnaire survey data, there are 1704 responses to the set of questions. The direct health-related questions are regarding symptoms, dis- orders, psychoses, depression attributes. For what can be the context, there are many other questions relating to: anti-social behaviour, eating characteristics, alcohol consumption, drug use, gender, age, educational level, marital status, region lived in, and employment status.

The initial analysis was to take as the active variables, the questions relating to neurotic symptoms and common mental disorders. For supplementary variables, comprising contextual variables, used were 9 socio-demographic variables. It was found that these socio-demographic variables, as supplementary variables, were mapped close to the origin. What is close to the origin, i.e. the zero value on the factor axis, is common and not at all special in its relationship with the factor. Figure 1 displays the six variables that contribute most to the factors, factors 1 and 2. This contribution is to the inertia that constitutes the factor. The following explains this principal factor plane.

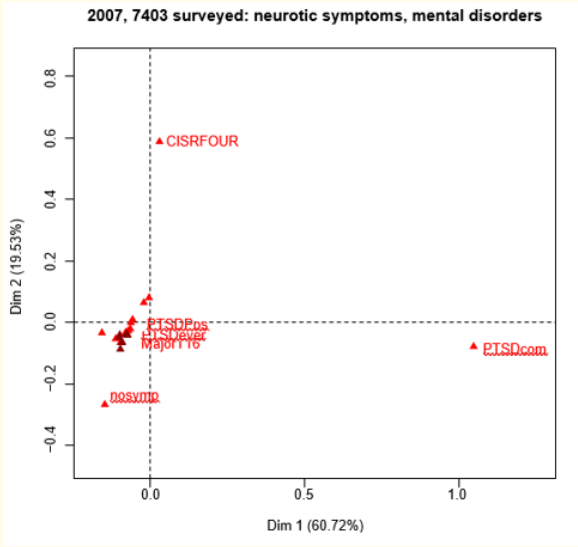

Figure 1: *Adult psychiatric morbidity survey 2007, England, household survey. The socio-demographic variables, as supplementary variables are close to the origin. Displayed are the 6 highest contributing variables.*

In figure 1, factor 1 is explained as the variable "PTSDcom", which are trauma responses, versus all other variables. Factor 2 is explained as the variable "CISRFOUR", being four common mental disorders, versus "nosymp", implying no symptom - no neurotic symptoms in the week prior to the survey carried out.

Since the socio-demographic variables were found to be close to the origin and therefore unrelated to the specific nature of the factors, therefore in the next step of the analysis, resulting in figure 2, the socio-demographic variables were the active variables to constitute the factors. Then the neurotic symptoms and common mental disorders questions were supplementary, mapped into the factor space, to be contextual here, for the given socio-demographic characteristics of the respondents. However, though here, what constitutes the supplementary, hence contextual variables were found not to be well located with their projections on the factors. That latter fact led to very limited usefulness here of the effect of these supplementary variables for interpretation.

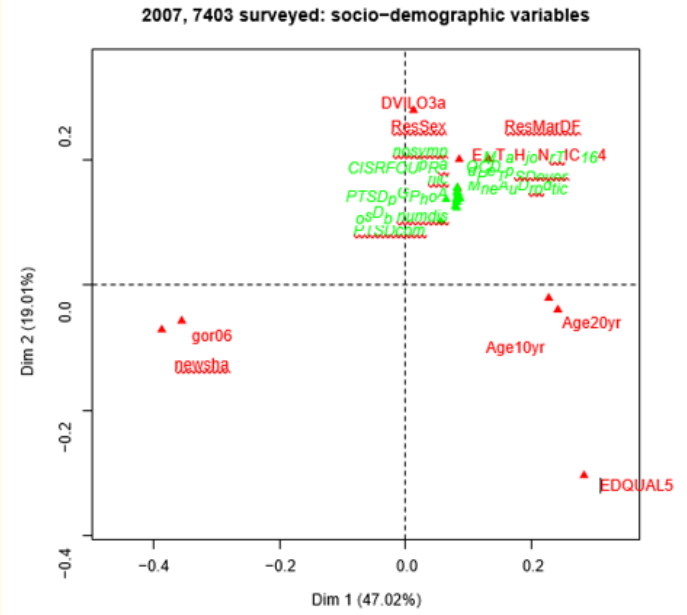

Figure 2: *Adult psychiatric morbidity survey 2007, England, household survey. The neurotic symptoms and common mental disorders, as supplementary variables (in green here) are close to the origin.*

In figure 2, factor 1 has age and education level counterposed to respondent's region location. In Murtagh and Farid [6], there is explanation for the variable names here. Factor 2 has educational level counterposed to: gender, marital status, employment status, and ethnicity.

Figure 1 used as the active variables for creating the factor space, neurotic symptoms and common mental disorders data, and socio-demographic supplementary, and therefore contextual data. Because very little was to be found from the latter, figure 2 used as the active variables for creating the factor space, the socio-demographic variables, and with the neurotic symptoms and common mental disorders data as supplementary variables and therefore con- textual. But there too, little was found from the latter. For other data sources, there can well be the need to carry out the analysis like it has been done here.

Following the two stages of analysis, with the principal factor planes in figures 1 and 2, we wanted to check the full set of data being analyzed together. This is neurotic symptoms and common mental disorders data jointly analyzed with the socio-demographic data. Figure 3 shows the principal factor plane for this outcome. In all cases here, the figures display highest contributing variables, the latter being questions and the contribution being to the inertia of the factors, in these displays the first and second factors.

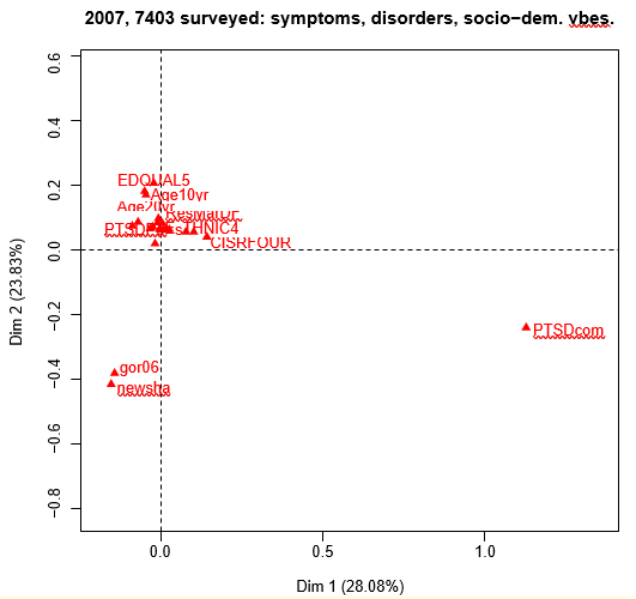

Figure 3: *Adult psychiatric morbidity survey 2007, England, household survey. The analysis has the neurotic symptoms and common mental disorders, and the socio-demographic variables. Displayed are the 10 highest contributing variables to the principal plane.*

In figure 3, on the positive first factor, there is a question relating to trauma ("PTSDcom"). The negative second factor has these two variables projected: "gor06", "newsha", and these are regional locations ("gor" here is "Government office region" and "sha" is "Strategic Health Authorities"). So, to briefly state the interpretation of this principal factor plane display, the first factor is interpreted as being recorded trauma, and the second factor is interpreted as being the regional location of the respondent.

Conclusion

Coenders, *et al.* [7] provide discussion of surveying, that can lead to relevant data sources.

Big Data inputs are required to calibrate and then with the potential to validate, by having comparability arising from various data sources. Open data sources are very likely to be key elements, and these may possibly arise from the technologies that produce relevant or important data sources.

The work at issue here was with: Adult Psychiatric Morbidity in England, a household survey, covering: Common mental disorders; post-traumatic stress disorder; suicidal thoughts, attempts and self-harm; psychosis; antisocial and borderline personality disorders; attention deficit and hyperactivity disorder; eating disorder; alcohol misuse and dependency; drug use and dependency; psychiatric comorbidity.

A brief description of analytical processes are as follows. Qualitative and quantitative observing and monitoring of wellbeing: new statistical drivers, Big Data analytics, Open Data, geometry and topology of data and information, semantics, homology and field.

Bibliography

1. Zhongheng Zhang F., *et al*. "Big Data Clinical Study and Its Implementation with R". AME Publishing Company, China (2018).
2. F Murtagh. "Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics". Chapman and Hall/CRC Press (2017).
3. N Keiding and TA Louis. "Perils and potentials of self-selected entry to epidemiological studies and surveys". *Journal of the Royal Statistical Society, Series A*, 179.2 (2016): 319-376.
4. F Lebaron. "How Bourdieu "quantified" Bourdieu: the geometric modelling of data". Chapter 2 in K. Robson and C. Sanders, Eds., Quantifying Theory: Pierre Bourdieu, Springer (2009): 11-29.
5. HSCIC, Health and Social Care Information Centre (National Health Service, UK). National Statistics Adult Psychiatric Morbidity in England 2007, Results of a household survey, Appendices and Glossary (2009): 174.
6. Murtagh F and Farid M. "Contextualizing Geometric Data Analysis and related Data Analytics: A virtual microscope for Big Data Analytics". *Journal of Interdisciplinary Methodologies and Issues in Science* 3 (2017).
7. Coenders G., *et al*. "Living conditions, interviewer effects and perceived well-being of the elderly. A Multiple Correspondence Analysis approach". In A Ferligoj and A Mrvar (Eds.), Developments in Social Science Methodology (2002): 125-146.

Volume 7 Issue 6 June 2019