Over 70 Million Blood Analyses in Nearly 600,000 French Adults between 2010 and 2015 Methodology of the Rubidium Big Data Study

Emmanuel Sorbets¹, François-Xavier Goudot¹, Nicolas Garnier², Nicolas Dubois³, Abel Betraoui², Vincent Bonnivard², Nathalie Charnaux⁴, Maud Schwoerer², Edith Martins-Meune⁵, Anna Szostek², Christophe Meune¹,6*

¹Department of Cardiology, Avicenne University hospital, APHP, Paris 13 University, Bobigny, France
²ProbaYes, Clinical Research Organization, Montbonnot, France
³Biorance Laboratoires Réunis, Saint Malo, France
⁴Department of Biochemistry, Avicenne University hospital, APHP, Paris 13 University, Bobigny, France
⁵Department of Geriatrics, Max Fourestier Hospital, Nanterre, France
⁶INSERM UMR S-942, Paris, France

Abstract

Objectives: Biological tests are key in the management of many patients, but very little is known about their current use. Big data analytics allow learning from our practice.

Methods: We conducted a registry from a real world dataset that included all analyses performed in a large subset of laboratories located in the French Brittany between February 17, 2010 and August 30, 2015.

We collected the following data: a) demographic characteristics, b) lists of tests ordered and prescription date, c) lists of prescribing physicians, and d) tests results.

Results: By September 2015, 22 laboratories were actively participating in the study, 7 in rural and 15 in urban areas. The aggregated data corresponded to raw data of 74,502,510 analyses performed in 585,745 distinct adult patients >20y. Male/female ratio was 0.73; mean age of the study population was 49.8y. Private laboratories performed 63% of the analyses, hospital-based laboratories performed 27% and 10% were performed by emergency medical facilities.

Conclusions: This study will examine the current use of biological tests, especially biomarkers, with the goal to better understand physicians’ knowledge of use of biomarkers, physicians’ prescriptions habits, as well as the determination of observed concentrations across different subgroups of patients.

Keywords: Big Data; Biomarkers; Medical Practice; Medical Registry.

Introduction

The choice of appropriate diagnostic procedures and effective therapies represents a regular challenge for most medical practitioners. Several biomarkers have recently been identified, which have become diagnostically, prognostically or therapeutically decisive [1]. While the clinical application of some biomarkers lowers the mortality [2], is endorsed by the latest practice guidelines [3,4] and may even be perceived as the “holy grail”, others remain the object of intense medical research [5].

Several well-known cardiovascular risk prediction models based on large study samples have been developed, including the Framingham Coronary Heart Disease risk score, the Thrombolysis in Myocardial Infarction and the CHADSVasc risk scores [6,7], which, however, are rarely used routinely, as the evidence is lacking that they markedly improve clinical outcomes [8]. More recently, a rapid increase in the generation of digital data combined with the development of mathematical applications have opened new perspectives from massive data sets and contributed to the current fascination from “big data” [9], a term variably defined, though generally referring to the 4Vs, i.e. Volume of data, Velocity of analysis, Variety of sources and Veracity, i.e. the trust that accurate data are available for their intended use [10,11]. The potential of big data analytics has been demonstrated in the development of predictive models and personalized medicine, surveillance of drugs and devices safety, the use of clinical decision support systems, population health analyses, and in research and educational messages [12].

Sources of medical big data abound, including administrative claim records, clinical registries, electronic health records, biometric data, the Internet, biomarkers and clinical imaging [13,14]. Theoretically, biomarkers are indicative of individual patient characteristics, and thus may allow improved medical management [2]. Another, thus far unreported, potential
application of biomarkers big data is the evaluation of medical procedures. This manuscript describes the methodology of a big data study, which focuses on the use and, perhaps, abuse of biological measurements (with a special focus on biomarkers) by caregivers, and the characteristics of a large sample population focusing on the concentrations of these biomarkers.

**Sample Population and Methods**

The defining use and misuse of Biomarkers Dosage and trends in observed concentration over time - a Big Data analysis (Rubidium) is a retrospective analysis of all tests [DNA and serology excluded] performed between the 17th of February 2010 and the 30th of August 2015 in laboratories from the Biorance group (Biorance Laboratoires Réunis, Saint Malo, France) [Table 1]. The Biorance group is a network of 22 laboratories and is associated with 11 additional laboratories, all located in the French Brittany. The 13,653 km² encompassed in this study cover most of the 6,775 km² Ille-et-Vilaine and part of the 6,878 km² Côtes d'Armor departments, and, in 2015, included a total of 4,568 physicians of whom 1,397 were general practitioners, representing respective densities of 327 and 247 physicians per 100,000 inhabitants, compared with 284 per 100,000 inhabitants in all of France.

All laboratories were equipped with Roche Diagnostics (Meylan, France) solutions, Elecsys® 2010 analyzers - cobas® 6000 analyzers - MODULAR ANALYTICS E analyzer and cobas e 411 analyzers. Within the scope of this study, the unit of observation represents the result of a single analysis, which could be transposed to the levels of individual patients, individual caregivers or all caregivers.

<table>
<thead>
<tr>
<th>Table 1: List of Tests Collected and Analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hematology</td>
</tr>
<tr>
<td>Electrolytes</td>
</tr>
<tr>
<td>Kidney function</td>
</tr>
<tr>
<td>Inflammation</td>
</tr>
<tr>
<td>Glucose analyses</td>
</tr>
<tr>
<td>Lipid analyses</td>
</tr>
<tr>
<td>Liver analyses</td>
</tr>
<tr>
<td>Thyroid analyses</td>
</tr>
<tr>
<td>Calcium-vit D analyses</td>
</tr>
<tr>
<td>Cardiac biomarkers</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Red blood cells, hemoglobin, platelet, white blood cells</td>
</tr>
<tr>
<td>Na, K, Cl, protein</td>
</tr>
<tr>
<td>Urea, creatinine, creatinine clearance (MDRD)</td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate, C-reactive protein</td>
</tr>
<tr>
<td>Glycaemia, HbA1C</td>
</tr>
<tr>
<td>Total-cholesterol, HDL-cholesterol, LDL-cholesterol, triglyceridemia</td>
</tr>
<tr>
<td>Alanine Aminotransferase (ALAT), Aspartate Aminotransferase (ASAT), Alkaline Phosphatase (ALP), Bilirubin, Gamma GT, Albumin</td>
</tr>
<tr>
<td>TSH, T3, T4</td>
</tr>
<tr>
<td>Calcemia, Vitamin D (total), 25-OH vit D, parathormon</td>
</tr>
<tr>
<td>Creatinine Phosphokinase (CPK), Troponin T, High-sensitivity Troponin T (Hs-cTnT), N-Terminal Pro Brain Natriuretic Peptide (NT-proBNP)</td>
</tr>
</tbody>
</table>

**Data collection and study sample**

We collected data applying to the general population of France and to the populations of the geographical areas under study from the following sources: 1) the “Institut National de la Statistique et des Etudes Economiques” (INSEE) [15] for the number and types of laboratories (rural versus urban, emergency versus elective), and the age and gender distribution of the study population, 2) the “Centre d’épidémiologie sur les causes médicales de décès » (CépiDc) for the mortality rates [16], and the “Conseil National de l’Ordre des Médecins” (French National Medical Council) for the number of physicians and their specialties [17].

We collected the following data in all patients >20 years of age who visited a participating laboratory during the study period:

a) Demographic characteristics,

b) Lists of tests ordered and prescription date,

c) Lists of prescribing physicians, and

d) Tests results.

All data were automatically collected by “middleware”, an application installed on dedicated computers in use by the participating laboratories, which enables data communications and management. Before being discarded, the blood samples were stored for one week after routine analyses, or longer after special analyses, as recommended by the French “Good Laboratory Practices”. No genetic analyses were performed. The data were extracted from middleware on September 10, 2015 and processed and aggregated in preparation for their analysis.

**Confidentiality protection**

At the first visit to one of 33 laboratories participating in this study, a unique, anonymous and fixed identification code was assigned to each patient. Data collected at that time included gender and date of birth. The results of the analysis were automatically assigned a barcode to an instrument data. Based on matched barcode identification, the instrument data were automatically populated with the patient date of birth and gender, test information, including list and dates of test requested and requesting caregivers, and date of blood collections and analyses. A fixed identification code was similarly assigned to each caregiver. For a subset of caregivers, the “Répertoire Partagé des Professionnels de Santé” (RPPS = shared registry of health professionals) number was disclosed, revealing their zip codes and specialties. After the latter were matched, the RPPS and zip codes were deleted. Ultimately, all data were anonymous and devoid of personal information, and pooled by age groups, gender, caregivers specialties, laboratory test name by specific patient.
samples. Finally, we examined the likelihood of a) finding a record that might reveal a patient’s identity, b) correlation, i.e. a link of ≥2 records from the same person or group, and c) inference, i.e. the ability to distinguish unequivocally a single value from a group of values. In absence of apparent risk of re-identification, we concluded that confidentiality was effectively protected [18].

Data management and analysis

Data cleansing and filtering: We began by rejecting ambiguous laboratory data, including calibration tests for the computer system, and the satellite laboratories, which sent only specialized analyses (e.g., toxicology, genetics) to Biorance, ultimately selecting 22 laboratories. We then excluded all analyses associated with 17 caregivers (out of a total of 22,965), whose identification codes were not composed according to the guidelines of the study protocol (identification code < 5 letters, exclusion has been made before anonymization process). Similarly, analyses from patients not referred by a known caregiver were not included in the study. Finally, all data associated with patients whose information was incongruous (E.g. age < 0 or > 119 years) were deleted. This filtering ultimately eliminated 756,917 (1%) of 75,259,427 analyses.

Treatment of missing data: A procedure dedicated to missing data was planned before the analyses. Missing demographic data were recovered, using other records from patients with the same identification code. For 1,724 out of 4,568 (37.7%) prescribers having ≥ 20 test orders, the RPPS number is known hereby allowing the record of the postal code, city location and specialty of the physician. No other data were missing after successful treatment of missing values.

Quality control: No patient was erroneously assigned both genders. Among the 376,169 patients who underwent >1 laboratory tests, the longest interval between two test orders was 5 years. Caregivers who, on average, ordered > 7.5 test orders per day (representing the 97.5th percentile), were flagged for being outliers. We considered that these overabundant requests for analyses were factitious and represented multiple medical institutions or offices sharing a single identifier. We examined the distribution of the values of each test and, when abnormal, the treatment of outliers on other variables.

Bias estimates: No estimate of biases planned before the analysis.

Ethical considerations

ProbaYes, a data science research organization in Montbonnot, France, as well as the 3 physicians François-Xavier Goudot, MD, Emmanuel Sorbets, MD and Christophe Meune, MD-PhD had exclusive access to the data and signed a privacy agreement with Roche Diagnostics, the study sponsor. The study protocol was approved by the ethics committee of Avicenne hospital, and the Commission Nationale de l’Informatique et des Libertés, the French agency regulating the protection of data, reviewed and approved the study.

Statistics

The initial analyses of the rUBIDuM registry’s aggregated data will 1) examine the characteristics of the biological tests ordered in a large sample population of patients and in various subgroups, 2) examine the results of the analyses, including blood chemistry, corpuscles and coagulation, cardiac enzymes, natriuretic peptides, lipid profile, and a special focus on biomarkers. Variations in results over time will be examined by comparing the yearly results between year 2011 and 2014. In case of major technical changes in the biological analyses or in the treatment, the analyses will be limited to fixed periods preceding the changes. All metric data will be treated by appropriate parametric (assuming underlining distribution) and non-parametric (distribution-free) statistics, without assumptions regarding the probability distribution of the variables compared. Some test results/measured concentrations are presented as categorical data (i.e. extreme values, text strings); these analyses were considered only to assess the number of test and test order but not in the calculation of the concentration.

Categorical data will be expressed as counts, percentages and ratios as well as cross tabulations (e.g. frequency table). All statistical tests will be performed appropriate to distributional requirements. Numerical variables will be compared, using Student’s t-test, Wilcoxon, Mann-Whitney U test or Analysis of Variance (ANOVA), as appropriate, and the chi-square test will be used to examine differences in frequency. All statistics will be performed based on the a priori hypothesis that the values are normally distributed. For multiple comparisons, the Bonferroni correction (adapt the p-value significance threshold) will be used, when appropriate. Due to the large sample size p-value < 0.001 will be considered significant.

All data aggregations and analyses will be performed by ProbaYes, using Python language and numeric computation libraries.

Results

By September 2015, 22 laboratories across French Brittany were retained for the study’s analysis. The studied population consisted of 74,502,510 analyses completed between February 17, 2010 and August 30, 2015, in 585,745 patients ≥20 years of age, of whom 209,576 (35.8%) underwent a single analysis. The mean of the total population of French Brittany is 41.3 years and that of all of France is 40.1. The mean and median ages of the 337,189 women and 248,556 men included in this analysis were 49.8 and 50.0 years respectively. These values are consistent with our inclusion of adults >20 years of age, in whom biological analyses were planned with the assumption that older are sicker than younger patients. Table 2 represents the number of analyses performed by gender and age-category.

Among the 3,294,302 inhabitants of Brittany 31,863 died in 2014 [16]. Based on the INSEE definition of the “tranche d’unité urbaine” (urban areas) which classifies a city of >30,000 inhabitants as an urban zone, 7 laboratories were in rural and
Table 2: Number of Test Order per Age and Gender

<table>
<thead>
<tr>
<th>Age Category</th>
<th>Test order (n)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>[10, 20)</td>
<td>34140</td>
<td>23901</td>
</tr>
<tr>
<td>[20, 30)</td>
<td>182105</td>
<td>50300</td>
</tr>
<tr>
<td>[30, 40)</td>
<td>334396</td>
<td>69458</td>
</tr>
<tr>
<td>[40, 50)</td>
<td>194127</td>
<td>110085</td>
</tr>
<tr>
<td>[50, 60)</td>
<td>200449</td>
<td>178799</td>
</tr>
<tr>
<td>[60, 70)</td>
<td>283893</td>
<td>323963</td>
</tr>
<tr>
<td>[70, 80)</td>
<td>292627</td>
<td>332314</td>
</tr>
<tr>
<td>[80, 90)</td>
<td>403299</td>
<td>345165</td>
</tr>
<tr>
<td>[90, 100)</td>
<td>197343</td>
<td>95578</td>
</tr>
<tr>
<td>[100, 110)</td>
<td>10075</td>
<td>2386</td>
</tr>
<tr>
<td>[110, 120)</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>2151634</td>
<td>1556065</td>
</tr>
</tbody>
</table>

Both gender 3707699

Figure 1 shows the crude death rates per 100,000 inhabitants in Brittany and in France, in men (1A) and women (1B). Among the 22 French regions, the minimum and maximum "Standardized Mortality Ratios" (the ratios of observed over expected deaths in the general population) were 189 in Île-de-France and 293 in the "North" region for men, and 116 and 183, respectively, for women. Respective Standard Mortality Ratios in French Brittany in men and women were 270 and 183.

The 74,502,510 analyses performed were from 3,707,699 orders written by caregivers. Caregivers who ordered the analyses were 80% general practitioners, 7% gynecologists and 2% cardiologists. The median number of analyses (interquartile range) per single order was 7 (1-15). During the study period, the median number of test order per patient during the study period was 2 (interquartile range 1-6), and ≥2 orders were written in 394,936 patients. Urban physicians ordered 63% and rural physicians 34% of the tests.
Discussion

The rUBIDIum study is a real world dataset registry developed with a view to examine the rate of biological test measurements ordered and analyzed in a large sample population from French Brittany, a >13,000 km² rural and urban territory. This offers a unique opportunity to examine important aspects of the patient health of a large French population over time, and prescription patterns of tests and test results by caregivers. We focused our attention to French Brittany as this region shares several characteristics such as age distribution, density of physicians and mortality rates with the general French population [16], represents both urban and rural populations, and is less susceptible to migratory variations than most other French regions [19].

The analyses that were performed were evenly distributed between private and public and between urban and rural medical laboratories, enabling multiple comparisons between types of facilities. Over the last decades, evidence-based medicine has emerged as the most reliable means of approaching the scientific truth, helping caregivers thereby in their quest for optimal decision making [20,21]. A key component of evidence-based medicine is the hierarchical classification of evidence. Rationales for this classification are the elimination of biases from the study and the guarantee that its findings can be replicated. Randomized trials are assigned the highest level of evidence, before systematic reviews of multiple observational studies, single observational studies and uncontrolled clinical observations. Since, however, randomized trials are not all designed or conducted properly, their results must be critically scrutinized [22,23]. Among the several potential flaws of clinical studies, an insufficient sample size is often a major limitation, precluding the detection of significant between-groups differences [9]. When performed in highly selected populations, randomized trials may not replicate the “real world” [24]. Analyses of a variety of big data, whose potential in healthcare are probably enormous and the subject of growing attention worldwide, may overcome these limitations [25].

In the last decades, biomarkers have emerged as key diagnostic, prognostic and therapeutic tools in the management of several diseases [1]. They have also been used in the risk stratification of large or very large samples of the general population, or of selected subgroups [26,27]. More recently, special attention was paid to the determination of the best algorithm for a dedicated biomarker (i.e. comparison between the 1h versus 3h algorithm for high-sensitivity troponins). Biomarkers progressively become not only valuable for rule out process (thanks to their high sensitivity), but could be useful for rule in process as well. We assume that in the future, biomarkers will become more and more essential and will even progressively replace clinical algorithm [28,29]. To our knowledge, the use or abuse of biomarkers in routine medical practice has not previously been the focus of a large study. This analysis of big data will enable a scrutiny of the use of biomarkers under various conditions, including rural versus urban environment, by family physicians versus specialists, and in various types of patients and disorders, and draw conclusions that may help in the formulation of recommendations and/or educational programs for their use.

While big data may overcome several limitations of usual statistical analyses, some concerns persist with respect to privacy and ethical issues. These include, but are not limited to, release of private patient health care data, inappropriate access to the use of patient data, and even the potential use of data to inappropriately profile patients and differentially provide care. In the Rubidium study, we provide special attention to ethics and privacy. Our study applied a very strict protection of privacy and confidentiality by the creation of two consecutive identification codes. Our final assessment revealed no instance of identity disclosure, correlation or data inference. We assume that all future studies should effectively ensure that confidentiality is effectively protected. The rUBIDIum study was also submitted (and validated) by our local ethics committees, by the CNIL (the French agency regulating the protection of data) and the protocol was registered. These processes may lead to reviewing process, comments and thus may reassure physicians and patients.

Limitations of our study

Our study design assumed that our sample population resembles a random sample of a European target population. This assumption is probably true overall, given the rigorous data cleansing and the quality of our procedures. Our sample size is large compared, thus minimizing the risk to reporting highly skewed results. Second, since we studied patients from French Brittany, our results might not apply to non-Caucasians. Third, our analysis is limited by our inability to collect much of the clinical data including diagnosis, treatment, or management costs. Fourth, since a single measurement was collected in over one third of patients, pairs of results could not be systematically compared. Fifth, among the 22 participating laboratories, data were collected between 2010 and 2015 in 18, between 2011 and 2015 in one, between 2012 and 2015 in one, between 2014 and 2015 in one laboratory and in 2011 for the last one. Sixth, while we collected data between years 2010 and 2015, yearly comparisons could only be made for the period between 2011 and 2014, during which the study population was complete.

Summing up, we still believe that our real world data observational study has a potential to generate valuable insights for French caregivers.

References

Over 70 Million Blood Analyses in Nearly 600,000 French Adults between 2010 and 2015 Methodology of the Rubidium Big Data Study

chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J. 2016;37:2129-2200. doi: 10.1093/eurheartj/ehw128


