Multidimensional Analysis of Fetal Growth Curves

Mario A. Bochicchio¹, Antonella Longo¹, Lucia Vaira¹, Antonio Malvasi² and Andrea Tinelli³

¹ Department of Engineering for Innovation, University of Salento, Lecce, Italy
{mario.bochicchio, antonella.longo, lucia.vaira}@unisalento.it
² Obstetric & Gynecology Department Santa Maria Hospital Bari, Italy
antoniomalvasi@gmail.com
³ Obstetric & Gynecology Department Vito Fazzi Hospital Lecce, Italy
andreatinelli@gmail.com

Abstract - Fetal biometry is considered the keystone in fetal well-being assessment. In particular, fetal growth curves are extensively adopted to track fetal sizes from the early phases of pregnancy up to delivery. In literature a large variety of reference charts are reported to consider the differences among different ethnic groups, but they are up to five decades old and they not address many other factors influencing the diagnose correctness. Therefore, they are rapidly becoming inadequate to support the melting pot of ethnic groups and lifestyles of our society. Starting from a detailed analysis of the limits of classical reference charts, the paper presents a new method, based on multidimensional analysis for creating personalized fetal growth curves. A simple implementation based on Open Source software and simulated data shows that the effectiveness of the proposed approach depends on the adoption of Big Data techniques on Cloud infrastructures.

Keywords - Cloud computing; fetal growth; personalized diagnosis; multidimensional analysis.

1 Introduction and Background

In obstetrics and gynecology, data coming for medical tests are often used for diagnostic and documentation purposes, giving the opportunity of a systematic data analysis to improve patient care and to develop new health-assessment techniques.

For example, throughout pregnancy ultrasound scans of the maternal abdomen are routinely used to track fetal growth and to assess fetal health. This is pursued by collecting fetal parameters such as Biparietal Diameter (BPD), Head Circumference (HC), Abdominal Circumference (AC), Femur Length (FL), Crown to Rump Length (CRL). To detect whether growth parameters lay within normal ranges, measured values are compared to statistical data (i.e. reference charts with fetal growth curves, showing average values of such parameters as a function of the gestational age). Abnormal fetal growth often indicates one or more fetal pathologies.
The importance of this test, proposed more than five decades ago by Lubchenco [11], Usher & McLean [21] and Babson & Brenda [1] is supported by a huge amount of scientific literature and by the clinical evidence, but the same literature clearly shows its main limitations:

— the number of patients considered in the studies (some thousandth) is low with respect to the total number of newborn per year (about 140 ML in 2012) in the World;

— the patients considered in the studies are not representative of the variety of anthropometrical factors due to ethnicity, to familial aspects and to other relevant internal and external factors;

— the commonly used growth curves are up to five decades old, they are not updated for the current population and they are not suitable to investigate temporal trends and dynamic aspects in fetal growth curves.

Indeed, fetal growth is influenced by a variety of factors, racial, social and economic among others, as well as specific medical conditions that may pre-exist or that may develop during pregnancy. Hence, it’s not surprising that fetal biometric parameters show a degree of variation from country to country and from area to area within the same country. Beyond ethnicity, many others factors affect fetal growth including fetus gender, physiological and pathological variables, maternal height and weight, drug or tobacco exposure, genetic syndromes, congenital anomalies and placental failure [8, 15, 16, 20]. Some authors addressed these problems providing an increasing number of fetal growth charts for specific groups and subgroups of population, but their studies suffer from a considerable methodological heterogeneity making them difficult to use for diagnostic purposes. Other authors, such as Gardosi [4] in 1992, proposed to adjust growth curves for most of the influential factors and introduced the idea of individualized fetal growth charts according to specific maternal and fetal characteristics. After about 20 years the interesting proposal, based on proprietary software and on a centralized application, did not produce documented results in the scientific literature.

In this scenario, from a theoretical point of view, the creation of personalized reference charts for fetal growth curves can be formulated as a multidimensional analysis problem on the biometric dataset routinely collected by doctors during pregnancy for fetal health assessment. Indeed, dimensions can be defined as the parameters (ethnicity, maternal sizes, familial aspects etc.) impacting the fetal growth. Moreover, each homogeneous (with respect to a given set of dimensions) group of patients can be considered as a subcube of the above mentioned biometric dataset. The multidimensional analysis problem can be also expressed in the form: “which is the normal range associated to the X biometric parameter of a Y-weeks old fetus belonging to the subcube defined by the Z dimensional parameters?”.

An important constraint, in order to design a system able to dynamically answer to these questions, comes from the problem size, which can be defined as a function of the number of newborn per year. In this case, considering the storage space needed for the biometric dataset and for the related data (order of magnitude of some PB), the distributed nature of the system (as detailed in Section 4) and the number of operations per newborn (order of magnitude
of millions or more) we can state that the system must be based on a suitable Cloud infrastructure and must adopt Big Data techniques to satisfy its requirements.

The paper is organized as follows: after defining the mathematical bases of the problem (Section 3), we will show how a conventional open source data warehouse (DW) system can be effectively used to compute the personalized fetal growth curves (Section 4). Then, we will show why DW systems are insufficient to effectively manage the whole dataset needed to supply the “Personalized Fetal Growth Curves” service on global scale, and how the problem can be solved by using Cloud infrastructures and Big Data techniques (Section 5). Section 6 is for conclusions and future works.

2 Related Works

2.1 Medical aspects

After the pioneering works of Lubchenco [11], Usher & McLean [21] and Babson & Brenda [1], more than five decades ago, fetal growth assessment is a well established and mature research field in obstetrics and gynecology [5, 2, 3].

The proliferation of further studies on specific subgroups of patients [23-27] and the related proposal of an ever increasing number of reference charts was characterized by a considerable methodological heterogeneity making them difficult to use for diagnostic purposes. As a consequence in clinical practice generic reference charts are preferred to specific ones, or to more complex approaches based on suitable mathematical models [26], because of their simplicity. To preserve the simplicity of the approach without loss of diagnostic power, some authors proposed the adoption of purposely developed software tools (Web Applications, Mobile Application,..) allowing to create customized growth charts [4,5,9] based on a regression model fitted to a very large group of newborns. On the other hand the World health Organization (WHO) standards are still based on generic reference charts, they don’t differentiate by ethnic origin and are not subject to frequent update, so they are unsuitable to assess the biometric parameters in several cases of practical interest.

2.2 Technical aspects

Despite widespread adoption of cloud-based solutions by most industries, cloud computing has been slowly embraced by healthcare and the biggest impeding factors are concerns about security and performance of cloud services causing the poor usage by the citizen (this is the official reason for the closure of Google Health and Revolution Health for example). Despite all opposition, many institutions have moved to the cloud to lower their storage costs and facilitate the exchange of images and medical record, such as Microsoft Health Vault [14], the Merge Healthcare’s Project Honeycomb [13] and the Accenture Medical Imaging Solution [22].


3 Fetal Growth Monitoring of Heterogeneous Populations

The detailed analysis of the fetal intrauterine growth monitoring is out of the scope of this paper but, in summary, the underlying idea is that: a) fetuses at the same gestational age, with similar genetic make-up (ethnicity, familial aspects, ...) and in similar situations (food, smoke, drugs, ...) are subject to a similar growth. This kind of fetuses will be referred, in the following of the paper, as Homogeneous Fetal Groups (HFG); b) a fetus with growth parameters different from that of its HFG is potentially pathologic. Monitoring is performed by measuring the length or the circumference of specific parts of the fetal skeleton (the above cited BPD, HC, AC, FL and CRL) by means of maternal trans-abdominal ultrasound pictures. For the purposes of this paper, for each given HFG at each gestational age (in weeks) the probability to find a given length can be approximated by a Gaussian distribution. This is justified by the natural anthropometric variability we observe in normal fetuses and by the errors associated to the measurement procedure. For diagnostic purposes we assume that fetuses are healthy if their sizes are between the 10th and the 90th percentile of the cumulative distribution function, they are pathologic if their sizes are out of the interval 3rd – 97th, otherwise they are borderline.

The problem outlined in Section 1 arises when we try to size the statistic sample for the evaluation of the Gaussian distribution: a small sample is not representative of anthropometric variability observed in nature while a large sample leads to a large variance. In both cases the resulting Gaussian is ineffective for diagnostic purposes. To solve this problem we can demonstrate that the right statistic sample with the smallest possible variance is the largest HGF as previously defined. Moreover, it’s easy to see that all HGF’s can be dynamically extracted from the whole dataset by means of multidimensional queries where fetal sizes are the measures [7] and the parameters (ethnicity, maternal weight and height, familial aspects, foods etc.) impacting the fetal growth are the dimensions. From the practical point of view, dimensional data can be obtained from the pregnant women by anamnesis and/or by clinical tests while fetal sizes are measured by means of ultrasound scans.

4 Data Collection and Analysis

For the purposes of this paper, fetal biometric data and maternal data are created by means of a simulator, but in a real scenario, for a fully functional system, we should consider that medical data comes from a number of heterogeneous sources (hospitals, EHR repositories, patients etc.) through different channels (web services, web applications, mobile applications etc.) in different format (HL7 or proprietary formats).

4.1 Problem size

Each year 140 millions of newborn come to the World (4.4 from US and 5.5 from Europe). Considering a Fetal Growth Tracking (FGT) online-service able to follow 50% of newborns from Europe and US, a record size of 10 KB for each pregnant
woman plus 2 KB for each fetus and a running window of 6 year to track mothers with 2 or more children, the global storage space sum up to about 360 GB of online multidimensional data (and about 1.5 PB including also ultrasound images). From the computational point of view, to generate the custom growth charts, the FGT system should run about 20 millions (5 millions pregnant women x 4 tests) multidimensional queries per year, i.e. about 2 queries each 3 seconds, which largely exceed the capabilities of a standard business intelligence suite (normally adopted for multidimensional analysis) such as Pentaho BI.

4.2 Data modeling and partitioning

For modeling the maternal and fetal data we started from the HL7 Reference Information Model [9] and from the Universal Data Model [18] which represent two good starting points in healthcare. Moreover, we analyzed the HL7 Common Data Model based on the Data Model Harmonization Process for the integration phase [9].

We assume that in the fully functional system data are sent to the system from a number of different data sources, most of which hosted in public or private clouds. The partitioning is due to the fact that each country is regulated by specific laws on health data management (privacy, security, accountability etc.) so that the granularity of data sources cannot be greater than that of the countries involved in the experiment. In general we can expect much lower granularity, from the hospital size to the single user size. Moreover, in several countries data protection laws require knowledge of where data are stored. This limited the adoption of the cloud in countries like Spain, France or Italy until cloud providers allowed obligatory data storage in a specific geographic location [12]. This aspect is important because, for further developments, we can imagine to exploit the distributed query capabilities typical of several modern clouds to index the ultrasound images and the other data and metadata produced for the fetal growth test and to retrieve it, according to specific criteria. This has the potential to improve the diagnostic process included in the fetal growth test and to support the inductive statistic approach, which is typical of Big Data research.

4.3 Multidimensional Analysis

In collaboration with a research group on obstetrics and gynecology we identified 8 main dimensions of analysis and 6 measures for the biometric fetal sizes. This lead to more than 10.000 Homogeneous Fetal Subgroups that the system must be able to identify and continuously update by adding the new fetal sizes to the system while they arrive from the hospitals or from the patients (20 ML per year, about 2 every 3 seconds). As shown in Section 5, this problem largely exceeds the computation capabilities of a classic multidimensional analysis system, such as the one included in the Pentaho suite. Advanced clustering techniques for multidimensional datasets based on MapReduce Framework [28] demonstrated the possibility to solve the problem at a reasonable price by means of Big Data techniques running on cloud infrastructure.
4.4 Privacy and Security

The issue of clinical data handling is well known in literature. Personal health records contain extremely sensitive information and thus pose strong privacy concerns. Preserving the privacy of medical data is not only an ethical but also a legal requirement, posed by several data sharing regulations and policies worldwide. Looking to the American world, for instance, such a project must comply with various regulatory policies (HIPAA, IPA, CMIA) and several security standards (ISO 17799, 27799, CD TS 21298, TS 21091:2005, TS 17090- 1:2002, 26000). This means that, since data come from the ultrasound equipment used for fetal health monitoring during the ambulatory visits, from anamnesis interviews and from other medical tests, special care has to be adopted to enforce data anonymity/privacy/de-identification. These issues are out of the scope of this paper, but in summary they have been addressed by separating the personal data from the clinical data using hash keys for reconciliation purposes.

5 Test Architecture

To have a first indication about the resources consumed by the FGT service and about how they scale with the problem size we created two dataset representative of the Italian demographic population (60 million inhabitants, 21 main ethnic groups). The two datasets, including respectively, 100,000 and 1,000,000 records, were produced by using the inverse Gaussian distribution function, generating respectively 500 and 5,000 Homogeneous Fetal Groups.

The 2 dataset were loaded on a Pentium 5 machine running at 2.5 GHz with 8 GB of RAM and 2 TB of storage space (SATA Disks). Then, they were processed with Pentaho 4.8 Community Edition, running 500 random queries to generate 500 different Personalized Fetal Growth Charts. The average elapsed times are shown in Fig. 3. They clearly demonstrate that even with a small fraction (about 3%) of the records to be processed in the fully working system, a single machine is far from satisfying the temporal constraints of processing 2 queries each 3 seconds defined in the previous section. As a second option we considered a recent evolution of the commercial version of Pentaho, called “Pentaho Business Analytics for Big Data”, which interfaces with Hadoop and other NoSQL and Analytic databases, permitting to create MapReduce functions and to visualize the results in Pentaho.
The detailed comparison between the results shown in Fig. 1 and results achieved with Hadoop will be completed in the next September and will be included in the final version of the paper, but they show a significant performance improvement, sufficient to meet the previously defined temporal constraint at a reasonable price. Two more options are under evaluations: a) the adoption of a different processing strategy, based on the clustering of multidimensional dataset with Hadoop, as described in [28]; b) the adoption of tensor-based computation, which is a very effective way to manage multidimensional datasets on distributed and parallel machines.

6 Conclusions and Future Works

The fetal growth assessment is a relevant problem since it concerns about 140 ML of newborns per year. Due to the population and “ethnicity reshuffling”, it’s by nature a global phenomena which can benefit from the adoption of Big Data techniques on cloud infrastructure. The goal is to obtain personalized Fetal Growth Curves’ computation within a given timeframe (< 1.5 s) over a given dataset (about 360 GB) of multidimensional data. In order to have an indication about the resources consumed by the FGT service we performed 2 different experiment based on Pentaho and Hadoop. The first results show that the problem can be solved at a reasonable cost and in an acceptable time. More promising results came from the adoption of 2 advanced approaches based parallel clustering algorithm and on tensor-based computation. In the future we plan to better explore these solutions comparing them with the Pentaho-Hadoop approach.

7 References


9. GROW software http://www.gestation.net


11. Lubchenco L. O., Hansman C., Boyd E. Intrauterine growth in length and head circumference as estimated from live births at gestational ages from 26 to 42 weeks. Pediatrics. 37, 403–408. 1966


17. PIC ULSS 8 Asolo available at https://www.ulssasolo.ven.it


