An estimation model for hypertension drug demand in retail pharmacies with the aid of big data analytics

Christos I. Papanagnou
Salford Business School
Salford, Manchester, M5 4WT
Email: c.papanagnou@salford.ac.uk

Omeiza Matthews-Amune
Salford Business School
Salford, Manchester, M5 4WT
Email: omeizam@yahoo.com

***Abstract*—The unpredictability of consumer preference observed in the last few years has coincided with the global digital data explosion as consumers are increasingly relying on internet information to guide their buying behaviour. The emergence of this trend has resulted in demand volatility and uncertainty in the retail industry, leading to negative consequences on inventory control and on shareholder profits in the long-run. This paper examines whether retail pharmacies in Abuja, Nigeria may exploit the increasing availability of relevant big data (structured, semi-structured and unstructured) from different sources to anticipate the changes on demand profiles for antihypertensive medication. In order to examine this, we consider a VARX model with non-structured data as exogenous to obtain the best estimation.**

***Index Terms*—big data analytics, hypertension, demand estimation, retail pharmacies**

I. Introduction

The concept of big data has evolved over recent years considering the rapid proliferation of data generated by humans, machines and nature in the form of digital processes in business, social media, videos, photos, sensors, mobile devices etc [1]. Ohlhorst (2012) described the term big data as the extremely large sets of data which have grown beyond the ability to be managed and analysed with traditional data processing tools [2]. Those companies that manage to take advantage of the power of big data are more likely to improve productivity and value creation by waste reduction leading to service improvement thereby providing the organization with a competitive advantage over peers [3]. Laney (2001) identified Volume, Variety and Velocity as the three major characteristics of big data [4]. However, other researchers have suggested an addition of Veracity and Value to the other three characteristics bringing it to a total of "5V's" ([5], [6]).

Given the several opportunities for big data (predictive) analytics, many companies strive to analyze enormous amounts of data to uncover underlying patterns which may contain hidden information that can assist in decision making. Multidisciplinary techniques, often referred to as data mining, can help companies to recognise and predict undesired phenomena [7]. In supply chain management such phenomena are often related to demand volatility and uncertainty, which affects demand planning and inventory holding. To minimize the impact of these challenges, many companies are increasingly investing in business intelligence in the form of text mining and analytics to leverage information from customer-generated content which can provide insights into consumer behaviour. The application of advanced analytics to the enhancement of visibility across the entire supply chain is referred to as supply chain analytics [8]. Supply chain analytics involves the application of quantitative techniques, which are often driven by technology to gain an understanding of the happenings within the entire supply chain so as to provide insights and improve prospective operational decision making [9]. The application of supply chain analytics to the analysis of complex data sets provides supply chain managers with the ability to respond to relevant problems in a timely manner by providing accurate business insights. Usually referred to as big data, such complex data sets which meet the 5 "V's" criteria could be unstructured (social media feeds, newspaper articles, emails, video etc.), semi-structured (weblogs, bar code data etc.) or structured (demand forecasts, ERP and CRM data etc.). Such unconventional data sources are increasingly supplementing traditional data sources such as sales data in the generation of short term business insight [8]. Utilising data analytics in strengthen supply chains increases the chances of timely customer order fulfilment subsequently ensuring brand loyalty and higher profit margins [10]. Data analytics could be applied across the supply chain at different levels namely forecasting, procurement, marketing, inventory management and traceability.

Companies consider different data sources and information to build commercial and customers' data profiles. Customer demand data may come from both internal (e.g., sales data) and external (e.g., social media) data sources and it may characterised by volume, velocity, and variety. In contrast to internal data sources, which they can be easily stored, accessed and analysed by companies, external data sources as not present within the companies can be collected by marketing, supply chain or IT intelligence departments usually

from huge user bases. However, linking external data to the overall company's commercial data environment is quite challenging as the use of intelligent tools that can measure the influence of exogenous variables is not yet a common practice. In terms of forecasting accuracy, there is a limited number of predictive analytics tools, which are based mainly on modelling data accumulated by business data sources. Current standard forecasting methods applied by businesses and discussed broadly in literature seem insufficient to elucidate the impact of external data sources [11]. Another characteristic of data sources is the structure of the data, which comes as structured (fixed format), unstructured (data without a fixed format) and semi-structured (a combination thereof) [1]. Verhoef *et al.* [6] argue that the synthesis of all those different data structures can help companies to understand better customer data on both supply and demand side. However, a big challenge for companies, nowadays, is to manage properly the sheer volume of unstructured data by assessing and utilising data mining through advanced processing systems like Apache Hadoop [12]. The objective of these systems is analysing and unearthing trends and correlations hidden within unstructured data, as they have the ability to improve business decisions and predictions thereby conferring competitive advantages on organisations. Also, some transmutation techniques allow the convertibility option into structured formats [13]. According to Gandomi and Haider [14], 80% of business data is classified as unstructured.

A. Demand estimation techniques in retail pharmacies

The need for an improvement in demand estimation in retail pharmacies, especially in developing countries, provides the motivation for this study as prediction accuracy within supply chains is pivotal to improving health outcomes. Results from Anusha *et al.*, (2014) indicate that among Indian pharmaceutical retailers, guestimates were the most common forecasting technique applied and such results which were usually unreliable often resulting in unavailability of life-saving commodities, which could be fatal or lead to the loss of customer loyalty and eventual profit reduction [15]. Studying sales trends among pharmaceutical distribution companies, Khalil *et al.*, (2014) identified prediction weaknesses as the major cause of excessive inventories and perverse drug shortages [16]. Retail sector of third world countries played an important role in the health care supply chain and inaccurate pharmaceutical forecasting often resulted in the unavailability of essential lifesaving commodities in some cases while in others, overestimation with wastages and expensive inventory holding as consequences [17]. The study also implicated forecasting errors in the loss of confidence within the health supply chains ultimately resulting in bullwhip effect. Yadav, (2015) also identified prediction challenges among pharmaceutical retailers as one of the causes of drug shortages in developing nations [18].

In the last two decades, several developing countries have experienced changes in their citizens' lifestyles and dramatic increase in average of their lives expectancies, and consequently, to rise of the adult population. Population growth alongside with lifestyle changes (such as lack of physical activity, tobacco use, behavioural risk factors and unhealthy diet) have been accused by researchers to constitute the main factors for the development of hypertension. The prevalence of hypertension is highest in the African Region as 46% of adults aged 25 and above diagnosed with hypertension [19]. In Nigeria, the prevalence of hypertension comprises a substantial portion of the total burden in Africa because of the large population of the country currently estimated to be over 170 million [20]. As this constitutes a very serious concern also across the globe, this paper examines whether retail pharmacies may exploit the increasing availability of relevant big data (structured, semi-structured and unstructured) from different sources to anticipate and predict changes on demand profiles for antihypertensive medication. To carry out this study, HMX pharmacy stores located in Abuja, Nigeria was used as a case study. Employing less than 100 people and classified as an SME, HMX operates as a pharmaceutical retail outlet providing a wide range of prescription only medications (POM) as well as other household consumables. Other services provided include doctor prescription fill-ups and pharmaceutical care. According to the literature, calcium channel blockers were the most widely prescribed antihypertensive medication by Nigerian doctors due to its high tolerability and affordability ([21], [22]). Guided by this, *Amlodipine* 10mg was selected because, from HMX sales data, it was identified as the most demanded calcium channel blocker in the last 5 years.

In terms of collecting data for this research, meeting the "5Vs" criteria are of critical importance. The Volume is represented by 30 months of data sourced from the pharmaceutical store, Google trends website, online newspapers and YouTube. The Variety is represented by the diverse data sources for this research and the Velocity is characterised by the speed of obtaining this data. The Value refers to the importance of these data in determining the impact of exogenous variables on forecasting accuracy while the Veracity signifies the reliability of the data obtained from trusted secondary sources.

II. Data sources and data types

In this research we consider data from prescription only medicines (POM). US Food and Drug Administration defines POM as medicines, which must be prescribed by a doctor and procured at a pharmacy for and intended to be used by one person through a signed prescription e.g., antihypertensive or antidiabetic medications [23]. Demand for these types of medication has been forecasted using quantitative techniques in the past. In forecasting demand for a POM used in treatment of hypertensive strokes, Bradford and Lastrapes argued that a simple method like the simple seasonal exponential smoothing outperformed more advanced ones like the AR(1) method in using a short data set to forecast the nonlinear consumption and the purchase of a drug [24]. Anusha *et al.* [15] compared different time series models (Simple Moving Average, Simple Exponential Smoothing and Winters Exponential Smoothing)

in forecasting POM and concluded that simple exponential smoothing performed better when forecasting demand in the Indian Pharmaceutical Retail company. In forecasting for prescription of antibiotics, Tesfamicael [25] compared ARIMA models, ESM (Exponential Smoothing Model), and the Heteroskedastic models (ARCH and GARCH), while in [26] a comparison analysis was conducted between a hybrid ARIMA-ANN, ARIMA and ANN models in forecasting demand for drugs used in hospital surgical operations. Gharbi *et al.* [27] also employed an ARIMA model in forecasting *Carbapenem* antibiotic resistance from antimicrobial consumption surveillance. In [28] several estimation-based techniques applied for use and costs of antineoplastic agents for the treatment of eye malignancies with the aid of a time series design with ARIMA (p, d, q) model.

A. Structured data

To carry out this study, weekly historical sales (demand) data from January 2014 to May 2016 collected from HMX pharmacy stores. The structured data for this study is generated by demand figures, which in this study are represented by the quantity of weekly sales of *Amlodipine* from 129 weeks as shown in 1. The use of medicine sales data has been broadly applied in the past mainly to predict demand patterns in the retail pharmaceutical industry (see [29], [26], [15]).

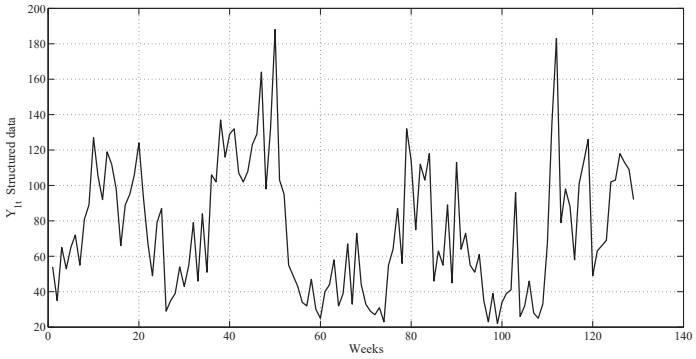

Fig. 1. $Y_1 t$, Structured demand Data Series for POM medicine

Visually examining the data plots presented in Figure 1 for *Amlodipine* medicine, the graph uncovers a cyclical pattern which Makridakis *et al.*, (2008) describe as rises and falls in the data plot that are not of a fixed period [30]. The unpredictability in demand could be because of the incidence of the ailment (hypertension) is also unpredictable and is not influenced by regularly occurring factors like other drug types such as analgesics but instead by physiological factors like hereditary conditions, body mass index, stress, coexisting conditions etc. The annual hypertension day which comes up on 17 May is marked with mass media campaigns and free screenings, which may be responsible for the increase in demand for both the medicines and information in the subsequent months.

B. Semi-structured data

Apart from well-formatted structured demand data, in our analysis we consider logical and systematic collection of semi-structured (meta)data from specific preexisting sources such as online newspapers (e-newspapers) and Google Trend websites. These types of (secondary) data do not require often recension as there is a certain degree of validity and reliability [31]. Given the option of predetermination and time-adjustment of the data window, there is a low level of inaccuracy or danger of analyzing outdated data. Semi-structured data is aggregated by considering the averages of the Google index (GI) and the Newspaper Keyword Index (NKI). The predictive power of searches in Google by using GI has been considerably investigated (see examples in [32] and in [33]). Sun *et al.* used GI and NKI to predict the fluctuation of the real estate market by combining online news articles and web search data [34]. Xu multiplied Google trend and Yahoo finance news data to obtain a unified index to forecast weekly stock price changes [35]. The keywords in this study are obtained from Google AdWords website, which provides the number of searched keywords relating to the indication of the medicine in the particular region. Based on search results, Table I presents the top ten searched keywords related to "hypertension" for POM medicines.

TABLE I
TOP TEN SEARCHED KEYWORDS FOR POM MEDICINE

Medicine type	Keywords
Hypertension	High blood pressure, stroke, hypertension, diabetes, cardiovascular diseases, heart disease, heart failure, heart attack, salt intake, obesity

1) Weekly Google Index: In order to collect the data representing Google searches for keywords related to hypertension, we have used the Google trends website, which stores weekly Google search data in terms of volume ratios of those keywords searched in specific geographical areas and shown in Table I. These ratios are calculated as a fraction of the total number of Google search queries for the same period. Then, the weekly GI is computed by aggregating the weekly Google trend values of all keywords had been searched for a particular area and then dividing each aggregated weekly values by the sum of all weekly figures for the examined period of study. To help the subsequent analysis, the values obtained are then multiplied by 100. The weekly GI, denoted as $Y_{2\_1(t,i)}$ is given in (1).

$$Y_{2\_1(t,i)} = \left[\frac{\frac{SK_{t,i}}{SQ_{t,i}}}{\max_t(\frac{SK_{t,i}}{SQ_{t,i}})} \right] \times 100, \quad t \in [1, T_w] \qquad (1)$$

where $SK_{t,i}$ represents the number of searches with keywords k for a given geographical area i at week t, and $SQ_{t,i}$ denotes the total number of search queries in geographical area i at each week t in total T_w weeks. It should be noted that the value of $\frac{SK_{t,i}}{SQ_{t,i}}$ is the keyword search ratio obtained

directly from Google trends. In case when Google searches occur in multiple geographical regions, $Y_{2\_1(t,i)}$ can be simply aggregated as it is given in (2), where r is the total number of different geographical regions.

$$Y_{2\_1(t,r)} = \sum_{i=1}^{r} Y_{2\_1(t,i)} \tag{2}$$

2) Weekly e-Newspaper Keyword Index: Results from a survey on Nigerian online newspaper reading perception showed that 79% of respondents read newspapers online [36]. Columbia University Libraries identified the 17 most read online newspapers in Nigeria, which are used for our research [37]. Weekly articles related to hypertension were selected for this research from these 17 online newspapers by Google queries. News content analysis for prediction has been applied by researchers in the past. A quantitative content analysis technique based on word count for specific keywords computation has been followed in [38]. However, the downside of the word count, is the use of synonyms, which may underestimate the importance of notions and multiple meanings and, consequently, may mislead the analysis. In order to avoid this barrier we may obtain the monthly scaled word by dividing the number of monthly occurrences of a specific (single) keyword by a proxy for the overall monthly text volume. In case more than one key topics are involved, a topic trend analysis was performed by averaging disease-related topic weights occurring in a social media blog over the same month to obtain a common monthly weighted average [11].

As we consider ten different keywords in this study, similar approach is applied to obtain first the "weight" of each keyword derived from Google *AdWords* used earlier, and to calculate then the unified weekly keyword index by averaging all keyword weekly weights. To obtain the individual keyword weight, all news articles from the selected e-newspapers containing the keywords in a single week are extracted. Afterwards, the frequency of each keyword, for each newspaper and for each day is divided by the total word count of the weekly aggregated articles, where keywords were found, to give a weight of the particular keywords during that week.

The weekly newspaper keyword index ($Y_{2\_2}(t)$) at week t is defined by (3).

$$Y_{2\_2}(t) = \frac{\sum_{i=1}^{N_d}\sum_{j=1}^{N_k}\sum_{n=1}^{N_n} f_i k_{j,n}}{\sum_{i=1}^{N_d}\sum_{n=1}^{N_n} TWC_{i,n}}, \quad t \in [1, N_w] \tag{3}$$

where $k_{j,n}$ denotes the j-th keyword at each newspaper n, and f_i represents the number of occurrences (frequency) at day i. $TWC_{i,n}$ denotes the aggregated total word count of articles at day i in each newspaper n, where each of the keyword k_j found. N_n, N_k, N_d and N_w represent the total number of newspapers (for this study, 17), keywords (we used 10 to facilitate our analysis), days and weeks, respectively. In order to ease our analysis we introduce the variable $Y_{2t} = Y_{2\_1(t,i)} + Y_{2\_2}(t)$ to measure the total semi-structured data. A time series plot of this data is shown in Figure 2.

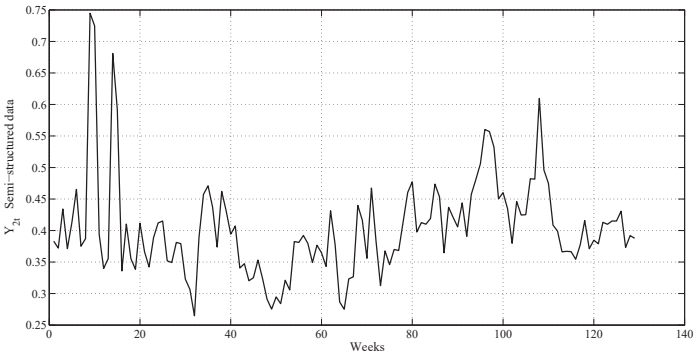

Fig. 2. Y_2t, Semi-structured demand data series for POM medicine

C. Unstructured Data

In this study, we consider unstructured data that is obtained from YouTube videos. YouTube is the third most visited website after Facebook and Google [39]. The utilisation of YouTube video data in determining the most popular and viral videos has been investigated by researchers (see [40], [41], [42]). Barfar and Padmanabhan (2015) employed conventional TV programme viewership data in predicting the outcomes of the 2012 US elections [43]. This approach can be also used to obtain the number of views per minute of viewers watching a specific YouTube video related to hypertension for a pre-specified period T. To obtain the number of views per minute, we amalgamate the weekly videos for a single keyword j and the total weekly video duration then is divided by the aggregated number of weekly YouTube views. The minutes per viewer $Y_3(t)$ at time t is obtained by (4) where TVD_t is the total video duration at week t and TNV_t is the total number of views at the same time period. Figure 3 provides a time series plot of unstructured data $Y_3(t)$. As it can be inferred from the plots in Figure 2 and Figure 3, some consistency is observed between the increase information seeking from mass media, Google and YouTube as shown by the slight spikes at the beginning of the years 2014, 2015 and 2016 also corresponding to increased drug demand spikes in the same period. However, there seems to be more consistency between the minutes per view of YouTube videos and POM demand shown in Figure 1. To ease the notation and analysis in the next section, we denote unstructured data as Y_{3t}.

$$Y_3(t) = \frac{\sum_j TVD_t}{\sum_j TNV_t}, \quad t \in [1, N_w], \quad 1 \le j \le 10 \tag{4}$$

Note that our approach can include higher volume sources of data (by simply encountering data from longer time windows), which can be processed by big data infrastructure components like Hadoop platform [12]. This technology can process and store massive amounts of distributed semi-structured and unstructured data derived mainly by online repositories

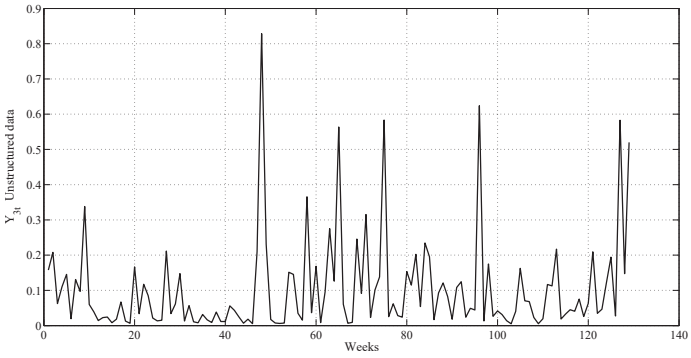

Fig. 3. Y_3t, Unstructured demand data series for POM medicine

by distributing it over multiple nodes in parallel [44]. Hadoop can also offer additional features such as scalability, speed, cost effectiveness, flexibility and resilience.

III. Model presentation and results

With the advent of Internet, more and more patients using the web-based applications to obtain health information [45]. The majority of health related Internet searches by patients are for specific medical conditions and medication. Most of them use a search engine while others search for broader terms (by looking for keywords and descriptions) rather than just typing the drug or brand name [46]. Since it is very likely of being referred to a doctor by a pharmacist (e.g., after having their blood pressure measured as high) or talk to a health care professional about the information they found, there is an indication that semi-structured and unstructured data can represent inputs to the actual demand figures (structured data). Thus, $X_t = (Y_{2t}, Y_{3t})'$ may be considered as a vector with exogenous variables to the system that influence the output - represented by structured data - Y_{1t}. However, initially in terms of model specification, we assume that the predictor variable $Y_t = (Y_{2t}, Y_{3t}, Y_{1t})'$. is a 3-dimensional vector time series for which an autoregressive moving average model needs to be specified. First, we attempted to estimate different VAR models of orders $m = 1, ..., 5$ to the 3-variable series Y_t by least-squares. The significance of each m-th order VAR matrix was tested by $H_0 : \Phi_m = 0$ against $\Phi_m \neq 0$, in case when a VAR(m) model has been fitted to the series. The tests were performed with the aid of likelihood-ratio (LR) testing principle by using the following statistics [47] $\lambda_m = (mk + 3/2 - T + m)\log(U_m)$, where $U_m = \frac{|S_m|}{|S_{m-1}|}$, $S_m = \tilde{\Sigma}_m \times T$ is the Maximum Likelihood (ML) estimator of an order m model and S_{m-1} is the sum of squared residuals for $\Phi_m = 0$. In order to identify the best order for the VAR model, we choose two selection criteria: Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (SBIC). The statistical results from fitting AR models to demand data are shown in Table II.

Results in Table II show the an AR model of order one or two can offer the best fitting among the other AR models. Since it is not clear which of two fittings provides better

TABLE II
Statistical results from fitting AR models to demand data for POM drug

m (VAR Order)	1	2	3	4	5
$\lvert\tilde{\Sigma}_m\rvert$	0.0588	0.0522	0.0464	0.0444	0.0414
λ_m	119.2031	14.2389	13.4763	5.0650	7.4209
AIC_m	-2.6520	-2.6695	-2.6411	-2.5391	-2.4583
$SBIC_m$	-2.6946	-2.6723	-2.6452	-2.5445	-2.4651

results, we will investigate both models. The LS estimates from the AR(1) model (with the corresponding standard errors of the ML parameter estimates in brackets) are:

$$\hat{\Phi}_{1,1} = \begin{bmatrix} 0.5548 & -0.0088 & -0.0002 \\ (0.0727) & (0.0427) & (0.0001) \\ -0.1541 & 0.0682 & 0.0001 \\ (0.1553) & (0.0912) & (0.0000) \\ -22.9485 & 2.6906 & 0.6495 \\ (30.9969) & (18.2109) & (0.0671) \end{bmatrix},$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.0041 & 0.0006 & 0.1962 \\ (0.0005) & (0.0000) & (0.1582) \\ 0.0006 & 0.0191 & -0.4508 \\ (0.0007) & (0.0023) & (0.3384) \\ 0.1962 & -0.4508 & 758.5120 \\ (0.1582) & (0.3384) & (94.8144) \end{bmatrix},$$

$$\hat{\alpha}_1 = \begin{bmatrix} 0.1980 \\ (0.0332) \\ 0.1452 \\ (0.0709) \\ 35.4081 \\ (14.1572) \end{bmatrix}$$

The LS estimates from the AR(2) model (with the corresponding standard errors of the ML parameter estimates in brackets) are:

$$\hat{\Phi}_{2,1} = \begin{bmatrix} 0.6380 & -0.0089 & -0.0002 \\ (0.0870 & (0.0422) & (0.0002) \\ -0.1486 & 0.0599 & 0.0002 \\ (0.1890) & (0.0918) & (0.0004) \\ -14.5209 & -1.2999 & 0.5019 \\ (6.8951) & (1.9113) & (0.0872) \end{bmatrix},$$

$$\hat{\Phi}_{2,2} = \begin{bmatrix} -0.1714 & -0.0586 & -0.0002 \\ (0.0868) & (0.0419) & (0.0002) \\ 0.0042 & 0.1209 & -0.0000 \\ (0.1885) & (0.0911) & (0.0005) \\ 3.3474 & 13.8644 & 0.2257 \\ (1.7907) & (1.7758) & (0.0879) \end{bmatrix},$$

$$\hat{\Sigma}_2 = \begin{bmatrix} 0.0040 & 0.0002 & 0.2416 \\ (0.0005) & (0.0008) & (0.1519) \\ 0.0002 & 0.0189 & -0.4493 \\ (0.0008) & (0.0024) & (0.3291) \\ 0.2416 & -0.4493 & 718.6423 \\ (0.1519) & (0.3291) & (90.1833) \end{bmatrix},$$

$$\hat{\alpha}_2 = \begin{bmatrix} 0.2466 \\ (0.0377) \\ 00.1282 \\ (00.0820) \\ 324.1509 \\ (15.9994) \end{bmatrix}$$

The results show clearly an one-way relationship between $X_t = (Y_{2t}, Y_{3t})'$ variables and output Y_{1t}, since the AR coefficient estimates in the $\hat{\Phi}(1,3)_j$ and $\hat{\Phi}(2,3)_j$ positions are close to zero and relatively small to their corresponding standard errors variables. This highlights the exogeneity of X_t to Y_{1t}. In contrast, the higher values on the lower left corner of each of the $\hat{\Phi}_j$ matrices denote that only structured data variables Y_{1t} depend on the past values of semi-structured Y_{2t} and unstructured data Y_{3t}.

Further, we can test our results by fitting two constrained models AR(1) and AR(2), by conditional ML, so that $\hat{\Phi}(1,3)_j = 0$ and $\hat{\Phi}(2,3)_j = 0$. This leads to the following conditional ML estimates of $\hat{\Sigma}_0$:

$$\hat{\Sigma}_{0,1} = \begin{bmatrix} 0.0042 & 0.0000 & 0.2010 \\ (0.0005) & (0.0008) & (0.1597) \\ 0.0000 & 0.0191 & -0.4546 \\ (0.0008) & (0.0024) & (0.3389) \\ 0.2010 & -0.4546 & 758.8261 \\ (0.1597) & (0.3389) & (94.8533) \end{bmatrix},$$

$$\hat{\Sigma}_{0,2} = \begin{bmatrix} 0.0041 & 0.0001 & 0.2504 \\ (0.0005) & (0.0008) & (0.1542) \\ 0.0001 & 0.0189 & -0.4549 \\ (0.0008) & (0.0024) & (0.3298) \\ 0.2504 & -0.4549 & 719.3229 \\ (0.1542) & (0.3298) & (90.2687) \end{bmatrix}$$

with corresponding selection criteria, $AIC_{0,1}$ =-2.6737, $SBIC_{0,1}$ =-2.6961 and $AIC_{0,2}$ = -2.6376, $SBIC_{0,2}$ = -2.6404, which favour constrained AR(1) model as the best for fitting. The final mode based on the LS estimates is given below:

$$\hat{\Phi}_{1,1} = \begin{bmatrix} 0.5599 & -0.0043 & 0 \\ (0.0733) & (0.0430) & (0) \\ -0.1581 & 0.0647 & 0 \\ (0.1554) & (0.0912) & (0) \\ -22.6161 & 2.9880 & 0.6650 \\ (31.0024) & (18.2134) & (0.0662) \end{bmatrix},$$

$$\hat{\alpha}_1 = \begin{bmatrix} 0.1781 \\ (0.0307) \\ 0.1608 \\ (0.0650) \\ 34.0927 \\ (14.1292) \end{bmatrix}$$

Also, the Hessian matrices for both unconstrained and constrained models were found positive-definite, since all eigenvalues were positive (for the constrained model $\lambda_{\min} = 2.74 \times 10^{-6}$ and $\lambda_{\max} = 8.99 \times 10^{3}$). Last, the relationship between structured data Y_{1t}, semi-structured Y_{2t} and unstructured data Y_{3t} in the form of a transfer function is given in (5), where B is the backshift operator. It can be inferred that the demand for POM medicine is influenced (in negative fashion) more from semi-structured data than from unstructured data.

$$(1-0.6650B)Y_{1t} = -22.6161BY_{2t} + 2.9880BY_{3t} + 34.0927 + \varepsilon_{1t} \quad (5)$$

IV. Conclusion

The challenge of demand volatility experienced by retail pharmacies and the available but limited capacity of estimation techniques have driven the need for this research to consider more data from exogenous sources that could indicate consumer behaviour and have been neglected in the past by researchers. The development of a multivariate time series analysis model, based on actual demand and consumers' generated content from a variety of internet sources, has the potential of providing estimation techniques and improving response to demand volatility in retail pharmacy.

This paper also highlighted the importance of business informatics intelligence through data extraction, mining and analytics, and it is believed that the results from this research will encourage further exploitation of big data predictive analytics, thereby proffering innovative solutions to tackle the challenge of demand uncertainty. Also, with the increasing government regulation in the sale of medicines and fierce competition by rivals - which continue to reduce business profits - this work underpinned the importance of incorporating exogenous user generated content. There are some limitations of this study, which we are aware of but we believe they do not impact significantly the results of this study. These are: (i) the low rate of Internet penetration in Nigeria (currently less than 50%), and, thus, the dependence on the secondary data obtained from the internet excludes observations from non-internet users, (ii) it is difficult to determine if videos have been watched to the end in order to have an impact on the viewers decision-making, and, (iii) the data analysis was restricted to the town of Abuja in Nigeria.

This study arguably focuses more on the variety characteristic of big data. It may be of value for future research to consider the volume characteristic by scaling up the data volume of the study to a minimum of five years. Also, in respect to the volume, in place of a single city as considered in this case, further research may also consider using country wide data for demand of other antihypertensive medicines and Google search intensity to determine if the results are consistent. The increase in the popularity of social media as a key source of user generated content also makes it imperative for social media data to be considered as an exogenous variable in future studies. Considering that demand volatility is a challenge across retail industries, it may be of note to consider

if the result from this study would be consistent if replicated in related sectors like food or other small-scale retail businesses. Last, the proposed multivariate model can be further developed to accommodate non-stationarity features and enhance the performance analysis by comparing estimation and predictive accuracy and forecasting with other known and widely-used univariate time series models.

REFERENCES

[1] J. Anuradha, "A brief introduction on big data 5 "V's" characteristics and Hadoop technology," *Procedia computer science*, vol. 48, pp. 319–324, 2015.

[2] F. J. Ohlhorst, *Big data analytics: turning big data into big money*. John Wiley & Sons, 2012.

[3] T. McGuire, J. Manyika, and M. Chui, "Why big data is the new competitive advantage," *Ivey Business Journal*, vol. 76, no. 4, pp. 1–4, 2012.

[4] D. Laney, "3-D data management: Controlling data volume, velocity and variety. META Group Inc. original research note," 2001.

[5] M. F. Uddin and N. Gupta, "Seven "V's" of big data understanding big data to extract value," in *Conference of the American Society for Engineering Education (ASEE Zone 1)*. IEEE, 2014, pp. 1–5.

[6] P. C. Verhoef, E. Kooge, and N. Walk, *Creating value with big data analytics: Making smarter marketing decisions*. Routledge, 2016.

[7] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[8] I. V. Rozados and B. Tjahjono, "Big data analytics in supply chain management: Trends and related research," in *6th Internafional Conference on Operafions and Supply Chain Management*, 2014.

[9] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.

[10] R. Zhao, Y. Liu, N. Zhang, and T. Huang, "An optimization model for green supply chain management by using a big data analytic approach," *Journal of Cleaner Production*, vol. 142, pp. 1085–1097, 2017.

[11] W. Kim, J. H. Won, S. Park, and J. Kang, "Demand forecasting models for medicines through wireless sensor networks data and topic trend analysis," *International Journal of Distributed Sensor Networks*, vol. 2015, p. 36, 2015.

[12] W. Dai and M. Bassiouni, "An improved task assignment scheme for hadoop running in the clouds," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, no. 1, p. 23, 2013.

[13] D. J. Power, "Using big data for analytics and decision support," *Journal of Decision Systems*, vol. 23, no. 2, pp. 222–228, 2014.

[14] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[15] S. L. Anusha, S. Alok, and S. Ashiff, "Demand forecasting for the Indian pharmaceutical retail: A case study," *Journal of Supply Chain Management Systems*, vol. 3, no. 2, 2014.

[16] N. Khalil Zadeh, M. M. Sepehri, and H. Farvaresh, "Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[17] U.S. Agency for International Development, "Deliver Project, Task Order 1. the logistics handbook: A practical guide for the supply chain management of health commodities. Arlington, VA," 2001.

[18] P. Yadav, "Health product supply chains in developing countries: Diagnosis of the root causes of underperformance and an agenda for reform," *Health Systems & Reform*, vol. 1, no. 2, pp. 142–154, 2015.

[19] World Health Organization, "A global brief on hypertension: silent killer, global public health crisis," 2013.

[20] J. T. Akinlua, R. Meakin, A. M. Umar, and N. Freemantle, "Current prevalence pattern of hypertension in nigeria: A systematic review," *PloS one*, vol. 10, no. 10, p. e0140021, 2015.

[21] A. D. Adedapo, W. A. Adedeji, A. M. Adeosun, J. Olaremi, and C. K. Okunlola, "Antihypertensive drug use and blood pressure control among in-patients with hypertension in a Nigerian tertiary healthcare centre," *International Journal of Basic & Clinical Pharmacology*, vol. 5, no. 3, pp. 696–701, 2016.

[22] J. O. Fadare, S. M. Agboola, O. A. Opeke, and R. A. Alabi, "Prescription pattern and prevalence of potentially inappropriate medications among elderly patients in a nigerian rural tertiary hospital," *Therapeutics and Clinical Risk Management*, vol. 6, pp. 115–20, 2013.

[23] United States of America Food and Drug Administration, "Prescription drugs and over the counter drugs: Questions and Answers, viewed on 20 February 2017, http://www.fda.gov/drugs/resourcesforyou/consumers/questionsanswers/ ucm100101.htm," 2015.

[24] W. D. Bradford and W. D. Lastrapes, "A prescription for unemployment? recessions and the demand for mental health drugs," *Health economics*, vol. 23, no. 11, pp. 1301–1325, 2014.

[25] M. A. Tesfamicael, "Forecasting prescription of medications and cost analysis using time series (Doctoral dissertation, University of Louisville), viewed on 03 February 2017, http://ir.library.louisville.edu/cgi/viewcontent.cgi?article=2423&context=etd," 2007.

[26] N. Riahi, S.-M. Hosseini-Motlagh, and B. Teimourpour, "A three-phase hybrid times series modeling framework for improved hospital inventory demand forecast," *International Journal of Hospital Research*, vol. 2, no. 3, pp. 133–142, 2013.

[27] M. Gharbi, L. Moore, M. Gilchrist, C. Thomas, K. Bamford, E. Brannigan, and A. Holmes, "Forecasting carbapenem resistance from antimicrobial consumption surveillance: Lessons learnt from an OXA-48-producing Klebsiella pneumoniae outbreak in a West London renal unit," *International journal of antimicrobial agents*, vol. 46, no. 2, pp. 150–156, 2015.

[28] J. C. Hsu, L. A. Gonzalez-Gonzalez, V. H. Lu, and C. Y. Lu, "Longitudinal trends in use of targeted therapies for treatment of malignant neoplasms of the eye: a population-based study in taiwan," *BMJ open*, vol. 6, no. 5, p. e010706, 2016.

[29] A. Ribeiro, I. Seruca, and N. Durỳo, "Sales prediction for a pharmaceutical distribution company: A data mining based approach," in *Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*. AISTI, 2016, pp. 1–7.

[30] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting methods and applications*. John Wiley & Sons, 2008.

[31] J. Tuhkuri, "ETLAnow: A model for forecasting with big data forecasting unemployment with Google searches in Europe," The Research Institute of the Finnish Economy, Tech. Rep., 2016.

[32] N. Askitas and K. F. Zimmermann, "Google econometrics and unemployment forecasting," *Applied Economics Quarterly*, vol. 55, no. 2, pp. 107–120, 2009.

[33] X. Li, J. Ma, S. Wang, and X. Zhang, "How does Google search affect trader positions and crude oil prices?" *Economic Modelling*, vol. 49, pp. 162–171, 2015.

[34] D. Sun, Y. Du, W. Xu, M. Zuo, C. Zhang, and J. Zhou, "Combining online news articles and web search to predict the fluctuation of real estate market in big data context," *Pacific Asia Journal of the Association for Information Systems*, vol. 6, no. 4, 2015.

[35] S. Y. Xu, "Stock price forecasting using information from Yahoo finance and Google trend," *UC Brekley*, 2014.

[36] O. W. Olley and S. T. Chile, "Readers' perception of Nigerian newspapers on the internet," *Journal of Philosophy, Culture and Religion*, vol. 4, pp. 26–34, 2015.

[37] Columbia University Libraries, "Electronic Newspapers of Africa by Country, viewed on 26 January 2017, http://library.columbia.edu/locations/global/virtual-libraries/african_studies/newspapers/country.html," 2017.

[38] K. A. Kholodilin, T. Thomas, and D. Ulbricht, "Do media data help to predict German industrial production?" *DIW Berlin Discussion Paper*, 2014.

[39] Ericsson consumer lab, "Internet Goes Mobile: Country report Nigeria. An Ericsson consumer insight summary report over-the-counter medications, viewed on 22 July 2016, https://www.ericsson.com/res/docs/2015/consumerlab/ericsson-consumerlab-internet-goes-mobile-nigeria.pdf," 2015.

[40] G. Fontanini, M. Bertini, and A. Del Bimbo, "Web video popularity prediction using sentiment and content visual features," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 289–292.

[41] F. Figueiredo, "On the prediction of popularity of trends and hits for user generated videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 741–746.

[42] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Performance Evaluation*, vol. 68, no. 11, pp. 1037–1055, 2011.

[43] A. Barfar and B. Padmanabhan, "Does television viewership predict presidential election outcomes?" *Big data*, vol. 3, no. 3, pp. 138–147, 2015.

[44] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*. IEEE, 2010, pp. 1–10.

[45] M. McMullan, "Patients using the internet to obtain health information: how this affects the patient–health professional relationship," *Patient education and counseling*, vol. 63, no. 1, pp. 24–28, 2006.

[46] G. Peterson, P. Aslani, and A. K. Williams, "How do consumers search for and appraise information on medicines on the internet? A qualitative study using focus groups," *Journal of Medical Internet Research*, vol. 5, no. 4, p. e33, 2003.

[47] G. C. Reinsel, *Elements of multivariate time series analysis*. Springer Science & Business Media, 2003.