Visualizing Data for Hypothesis Generation Using Large-Volume Claims Data

Eberechukwu Onukwugha, PhD, University of Maryland School of Pharmacy, Baltimore, MD, USA; **Margret Bjarnadottir, PhD**, University of Maryland College Park, College Park, MD, USA; **Shujia Zhou, PhD**, University of Maryland Baltimore County, Baltimore, MD, USA; and **David Czerwinski, PhD**, San José State University, San Jose, CA, USA

Eberechukwu Onukwugha, PhD

KEY POINTS . . .

With technological advances it has become important to understand the role for and the potential of hypothesis-generating analysis *vis-à-vis* informed data exploration.

This article describes the application of visual analytics and a statistical, data-driven algorithm (i.e., Grouping Algorithm for Cancer Data) to claims data.

Visual analytics and data-driven algorithms provide intuitive visual representation to explore data, identify systematic patterns (including unexpected patterns), and unlock insights that are not typically possible through traditional statistical analysis.

This is the first of two articles in this issue on the topic of algorithmic advances in HEOR. Dr. Onukwugha et al. illustrate the potential for insight and hypothesis generation from observational data via the use of EventFlow and the Grouping Algorithm for Cancer Data.

Introduction

Hypothesis-driven data analysis is the traditional cornerstone of the advancement of medical knowledge. These analyses will typically rely on informed, carefully-developed hypotheses generated from prior studies. Exclusive reliance on statistical hypothesis testing for evidence generation may result in missed opportunities to develop new evidence from observational data, as statistical hypothesis testing is designed to answer previously-formulated questions and not to identify new areas for inquiry. As the growth in health information technology increases the availability of big data (i.e., large-volume, high velocity and varied data) for health services research, we have the opportunity to consider new approaches to generate evidence from observational data, including visualization tools and data-driven algorithms. These exploratory methods include both data-driven statistical analysis and visualization. We briefly distinguish hypothesis testing from hypothesis generation and then illustrate the latter with two case studies.

Hypothesis testing (HT) research utilizes existing bodies of knowledge and data to specify an *a priori* hypothesis, which is tested through experiments designed to produce relevant data which can be further interpreted [1]. This method is useful to investigate the relationship between a dependent variable and independent variables and has been traditionally used as the primary research study design to solve scientific problems. Simply stated, HT starts with an idea and uses data to empirically confirm or reject a hypothesis.

HT is useful when the researcher knows which parameter and variables are desired to study; however, when variables are subjective or no meaningful hypothesis can be stated, hypothesis generating (HG) designs [2] can be of value. In this context, the HG design can be employed to explore relationships in the data, to produce insights and to develop potential hypotheses for further study. Hypothesis-generating research analyzes data searching for relationships and patterns, and then proposes an explanatory hypothesis based on the findings [3]. The hypothesis may then be tested, to either refute or support the theory, in subsequent clinical research evaluations on datasets different from those used to develop the insights. This step (referred to as out-of-sample testing) is vital, as one of the limitations of HG exercises are that insights may be coincidences in the dataset; patterns that may seem compelling may be random rather than based on substantive information. HG does not have the constraint of a specific hypothesis, which allows this method to explore multiple outcomes and pathways. In sum, HT and HG are distinct study designs and have a role in evidence generation.

Recent advances in statistics, machine learning, and visualization enable researchers to go beyond traditional summary statistics and interactively explore large datasets to generate insight and develop hypotheses. Traditional hypothesis testing focuses on group differences in ‘1st order measures’ (e.g., averages), which, due to the large sample sizes in large-volume claims data, are typically highly statistically significant and therefore less useful for distinguishing systematic patterns. Specialized visualization tools allow drill-down into data sets and the exploration of group differences in ‘2nd order measures,’ such as “time between events” and “order of events.” Specialized tools compare two cohorts and automatically create possible hypotheses for exploration by the researcher [4]. Furthermore, advances in machine learning, including computationally effective clustering, allow us to derive new insights that are not possible with traditional analytical approaches [5].

In order to illustrate the utility of these advanced analyses for hypothesis generation, we first investigate the role of visualization on claims data including a specific visual analytics tool (VAT) for pattern summarization. Second we analyze the application of clustering to claims data. Both case studies demonstrate the value of intuitive visual representation to explore data, identify systematic patterns (including unexpected patterns), and unlock insights that are not typically possible through traditional statistical analysis. The first example illustrates the use of a VAT EventFlow [6,7], to explore prescription patterns [8]. The second example illustrates the use of a grouping algorithm[9,10] to investigate survival and cost accumulation.

Case Study 1 – Visualization of Hypertensive Prescriptions and Medical Claims Data

When analyzing claims data, visualizing sequences of events can be more revealing than studying summary statistics, and can help guide data cleaning, develop new insights and lead to hypothesis generation. Visualizations can for example focus on the duration of events, the order of events, and time between events, allowing for analysis of some of the 2nd order statistics described above.

To demonstrate the use of visualization for data exploration, we use a simplified example of a hypertensive population and their claims data that contains hypertensive prescriptions. We can view each member and his/her prescriptions and medical visits as a sequence of events. Prescriptions are viewed as interval events (i.e., from the day the prescription is filled until the end of the supply). In this example, medical visits are viewed as point events.

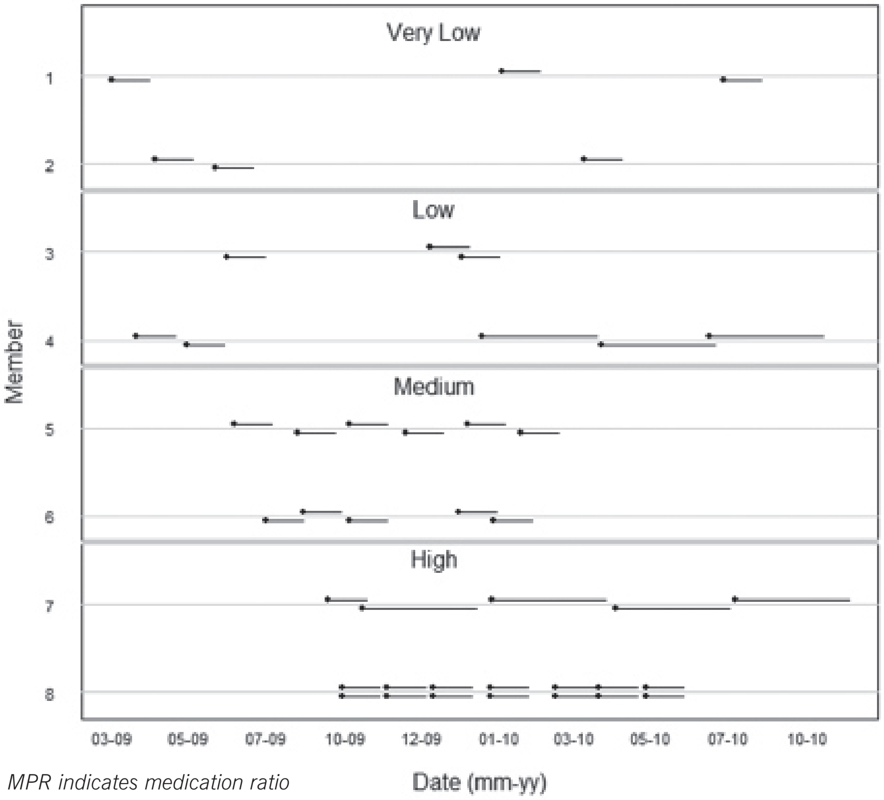
Among patients prescribed hypertensive medications, adherence is measured by the medication possession ratio (MPR) which can vary substantially between patients. While calculations of MPR provide a quantitative measure of adherence, and average adherence of a population is a widely-used quantitative measure of population adherence, visualization provides a unique perspective on the prescription filling patterns of patients. This unique viewpoint can spark hypothesis generation related to prescription fill patterns. Once patterns are identified visually, with careful thought they can then be defined quantitatively for subsequent out-of-sample hypothesis testing. Further, visualization can aid in identifying missed steps in data cleaning and/or help guide this stage of the analysis.

Figure 1 is an example of eight hypertension patients drawn from a large database. Two representative members are shown from four strata of MPR: very low (0 to <0.25), low (0.25 to 0.5), medium (0.5 to <0.75), and high (0.75 and above). These members were identified by selecting random samples from a large data set. Random sampling can be an effective method for examining overall patterns in a large dataset. Patterns that are prevalent in the dataset will also appear in a sample, and a sample can be easier to process visually.

Figure 1 reveals interesting patterns both within and between MPR strata. Among patients with very low MPRs, the visualization shows that months typically go by between prescription fills. Member 2 filled two prescriptions before a long gap occurred, while member 1 never had two successive fills. The review of these patterns can guide follow-up work to identify commonalities or distinguishing features among a larger sample and/or focused on individuals with very low MPR. In the second panel, showing patients with low MPRs, there are also large gaps between fills but with more episodes of continuity. By the end of the data window, member 4 had three successive 90-day fills and appears to be adherent. Member 3 did not have another fill after January 2010. Without the visualization, the similarity of the MPR values of these two members would have hidden their strikingly-different behaviors.

Among the members with medium and high MPRs, different patterns emerge. Prescriptions are filled much more consistently, with smaller gaps. Member 8 has two prescription fills on each fill date. This could be erroneous duplication in the data and should be investigated further and possibly cleaned up.

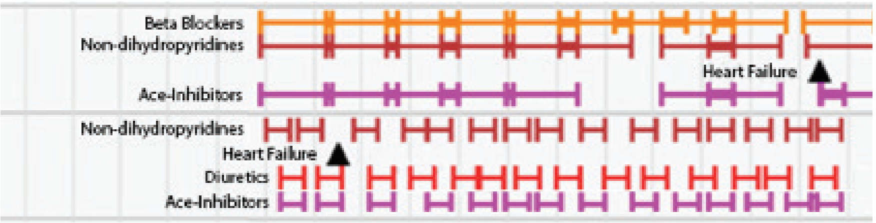
Figure 1. Visualization of hypertensive drug claims for eight members, stratified by MPR. Drug prescriptions are represented as intervals, each interval representing the duration of a single prescription.



MPR indicates medication ratio

>

Figure 3. Visualization of hypertensive drug claims and cardiac events for three members.



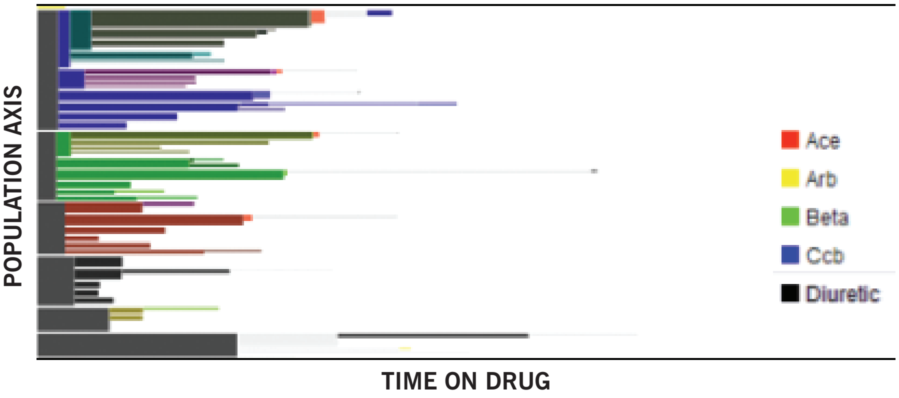
Drug prescriptions are noted as an interval, each interval representing a single prescription. Inpatient stays associated with heart failure are noted by black triangles.

By visualizing members with a range of MPR we were able to gain a more nuanced understanding of the variety of prescription filling behaviors. We also detected possibly erroneous data that indicates more data cleaning could be needed before proceeding with the analysis. The diversity of refill patterns, if confirmed in a larger random sample, could be used to inform decisions about whether to aggregate or further sub-divide the MPR categories or to investigate whether the overall patterns are preserved across subgroups of interest (e.g., age, race, comorbidity status). Furthermore, these visualizations can guide the researcher in drilling down in the claims data to provide clues to the underlying reasons why different members have different refill behavior.

The visualization in Figure 1 used random sampling to sample prescription patterns. However random sampling is not an effective technique for visually detecting rare patterns, since if a pattern is rare enough it is unlikely to be present in any given random sample. Further, random sampling does not efficiently summarize patterns among subgroups present in large populations. Software tools exist to visualize large data sets in such a way as to bring order to a large sample and make it more amenable to visual interpretation. Such organization can come from sorting, grouping, and other more sophisticated techniques. Figure 2 presents an example summary generated by a VAT called EventFlow [7].

Figure 2 gives an overview of the prescription patterns of members starting on diuretics. From the figure, it is evident that in most cases another drug class (e.g., CCBs) gets added, often after a short time. Further analysis and investigation can provide insight to better understand differences among the groups of patients who switch from diuretics to different classes. Coupled with data on outcomes (e.g., hospitalization), these visual summaries can generate hypotheses for further investigation. For example, Figure 3 is a visualization of three hypertensive patients that are switched away from non-dihydropyridines (a subclass of CCBs), to other drug classes. The figure includes heart failure-related hospitalizations as point events (indicated by triangles). For patient 1 there are multiple prescriptions until the patient is switched away from non-dihydropyridines following the hospitalization event, and for patient 3, the switch is not observed. Follow-up analyses could utilize a larger sample of individuals with prescriptions for CCBs and visualize non-dihydropyridine and dihydropyridine prescribing patterns separately for those with and without a heart failure-related hospitalization. Further statistical analyses could then characterize the use of non-dihydropyridines among heart failure patients and test hypotheses about the relationship between non-dihydropyridine use and heart failure-related hospitalizations.

Figure 2. Prescription patterns of members starting on diuretics that have filled at least one prescription for each of the five hypertensive drug classes.



Ace indicates angiotension converting enzyme; Arb angiotension receptor beta blocker; Ceb calcium channel blocker.

In addition, EventFlow and other visualization tools can support data-driven analysis by demonstrating the impact of modeling decisions (e.g., allowable prescription gap) on outcomes measures such as cost and adherence [8].

Case Study 2 – Application of a Grouping Algorithm to Cancer data

Machine learning algorithms offer an advanced way to view data, learn about systematic patterns in the data, or develop hypotheses for testing. As an example, the Grouping Algorithm for Cancer Data (GACD) [9] was developed with large sample sizes in mind, emphasizing computational time and accuracy, and is particularly efficient when processing data with multiple factors. The algorithm can provide unique information for the researcher with regards to the outcome of interest, including clinical events (e.g., survival) or costs. The GACD utilizes quantifiable measures (i.e., factors) to first group individuals into mutually exclusive groups based on their characteristics, and then in a second step groups the homogenous group of patients together into super-groups (or clusters) that are similar in terms of survival. The number of super-groups or clusters is chosen by the analyst. Table 1 provides an example of a group of homogenous patients that could be formed from the following five measures: cancer stage, age, race, prior hospitalization, and geographic location:

This example group represents individuals diagnosed with incident stage 4 cancer, aged 75 to 79 at diagnosis, African-American race, without evidence of

Table 1: Example of a natural cluster or 'combination'

Factors	Value
Stage	Stage IV
Age	75-79
Race	African-American
Hospitalization in the 12 months prior to cancer diagnosis	Not relevant indicator
Geographic location	Urban

hospitalization during the 12 months prior to diagnosis, and residing in an urban location at the time of diagnosis. The number of combinations increases exponentially with the number of factors available for use. The grouping is accomplished in four steps, and we refer the interested reader to published work [9] for the details. Figure 4 (a) represents a choice of five super-clusters, resulting in five survival curves. Group 3 (black) represents individuals with the best prognosis while Group 4 (yellow) represents individuals with the worst prognosis. Costs curves [10] associated with the individual survival curves can be created and an example curve is presented in Figure 4 (b).

The characteristics of the groups themselves as well as comparisons across groups can support hypothesis generation. For example, when we applied the GACD to cancer data (9), we found interesting patterns: 1) the group with the poorest survival was composed of white, non-Hispanic individuals living in urban areas, with multiple comorbidities at diagnosis and diagnosed at age 75 or older; 2) the rank ordering of groups based on the Charlson Comorbidity Index changed between the 12 month pre-diagnosis and 12 month post-diagnosis period, suggesting shifts in comorbidity burden following cancer diagnosis; 3) particular groups stood out in terms of the rate of increase in the cost curves at the end of the time period. Designed studies using a different sample could then build on these exploratory findings to further investigate prognostic factors among subgroups, the impact of longitudinal changes in comorbidity burden, and cost drivers.

Generating evidence for real-world impact

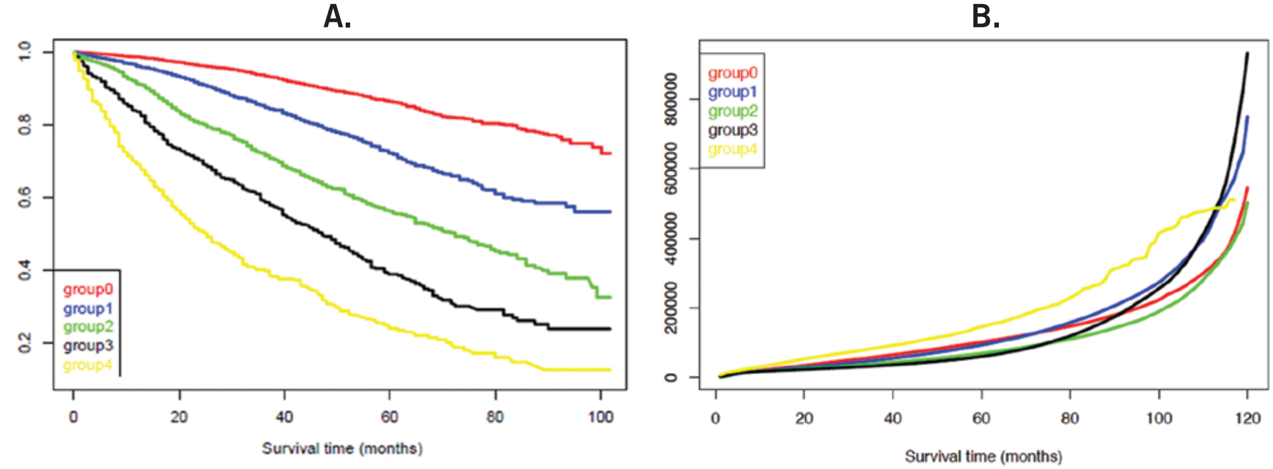
The examples above illustrate paths to hypothesis generation using visual analytics tools and a grouping algorithm. The use of these approaches will not be applicable, feasible, or even appropriate for all situations. The analyses described in the case studies require a comfort level with visualizing data, trusting the interactive, iterative process that is required to generate insight, and using secondary data for exploratory analysis. Combining the visual output with text summaries, the researcher can document the entire process, including all the alternatives considered along the way. With the documentation of the process elements (e.g., data, measures, base and alternative scenarios), the hypothesis-generation process can be subjected to review, critique, and improvement where appropriate.

The analyses described in the case studies are ideally suited for longitudinal data. This focus on the timing of events and event sequences provides opportunities to investigate the timing of various events that are of interest to multiple stakeholders. The studies can incorporate contextual (e.g., geographic location, area-level socioeconomic status, physician-specific, institution-specific) information, as 'attributes' in EventFlow and as factors in the GACD. The incorporation of contextual factors allows one to consider location-specific prediction models of events and associated costs.

In the past few years, computing power available to health care researchers and practitioners has been increasing dramatically due to the growth of commodity-based computer cluster, surging cloud computing services, as well as the ever-growing and maturing Hadoop/Spark ecosystem. This has enabled analyzing large amounts of data (e.g., over a Terabyte) with sophisticated methods from statistics and machine learning. Those advances offer promising opportunities to store and share a variety of data (e.g., electronic medical record systems, hospital business/operations records) and perform sophisticated studies. In particular, this kind of interoperability provides the opportunity to develop richer datasets which increases the number of different factors that could be used in the analyses.

In summary, we have illustrated example paths from visualization to hypothesis generation using health services research studies [8,10]. The analyses described

Figure 4: (a) shows the survival curves associated with 5 clusters and (b) shows the corresponding cumulative cost curves



>

above are appropriate when there is an interest to study the behavior of interval and point-based measures over time; can be used across various disease settings to summarize data and; provide graphical output, some of which may be unexpected, intriguing, and worthy of further exploration in designed studies. While we focus here on two case studies, there are many more examples [11] that are available for data visualization and interactive data analysis. Use of these tools for hypothesis generation can provide transparency and facilitate review, critique, and discussion of this important but sometimes misconstrued stage of research. When the analytic objectives and measures are appropriate, data visualization offers numerous possibilities for insight and hypothesis-generation using large-volume health care data.

References

[1] Hartwick J, Barki H. Research Report—Hypothesis testing and hypothesis generating research: an example from the user participation literature. Inf Sys Res. 1994;5:446-449. [2] Auerbach C, Silverstein LB. Qualitative Data: An Introduction to Coding and Analysis: New York, NYU Press; 2003. [3] Biesecker LG. Hypothesis-generating research and predictive medicine. Genome Res. 2013;23:1051-1053. [4] Malik S, Shneiderman B, Du F, et al. High-volume hypothesis testing: systematic exploration of event sequence comparisons. ACM transactions on interactive intelligent systems (TiiS) 2016;6(1, Article No. 9). [5] Bertsimas D, Bjarnadottir MV, Kane MA, et al. Algorithmic prediction of health care costs. Operations Res. 2008;56:1382-1392. [6] Monroe M, Lan R, Lee H, et al. Temporal event sequence simplification. IEEE Trans Vis Comput Graph. 2013;19:2227-2236. [7] Malik S, Du F, Plaisant C, et al. EventFlow. Software. Available at http://hcil.umd.edu/eventflow/. [8] Bjarnadottir MV, Malik S, Onukwugha E, et al. Understanding adherence and prescription patterns using large-scale claims data. Pharmacoeconomics. 2016;34:169-179. [9] Qi R, Zhou S. Simulated annealing partitioning: an algorithm for optimizing grouping in cancer data. IEEE 13th International Conference on Data Mining Workshops; December 7-10, 2013; Dallas, TX2013. [10] Onukwugha E, Qi R, Jayasekera J, Zhou S. Cost prediction using a survival grouping algorithm: an application to incident prostate cancer cases. Pharmacoeconomics. 2016;34:207-216. [11] Onukwugha E, Plaisant C, Shneiderman B. Data visualization tools for investigating health services utilization among cancer patients. In: Hesse BW, Ahern D, Beckjord E, eds. Oncology Informatics: Using Health Information Technology to Improve Processes and Outcomes in Cancer: Academic Press; 2016.

Additional information:

The preceding article is based on a workshop given at the ISPOR 21st Annual International Meeting.

To view the authors' presentations, go to: http://www.ispor.org/Event/GetReleasedPresentation/686; http://www.ispor.org/Event/GetReleasedPresentation/687