# Multivariate data-driven decision guidance for clinical scientists.

29th International Conference on Data Engineering Workshops, ICDEW 2013;
Proceedings - International Conference on Data Engineering

Published version available at
http://doi.org/10.1109/ICDEW.2013.6547449

# Multivariate Data-Driven Decision Guidance for Clinical Scientists

Frada Burstein[1], Daswin De Silva[1], Herbert F. Jelinek[2], Andrew Stranieri[3]

[1]*Centre for Organisational and Social Informatics,*

*Faculty of IT, Monash University, Victoria, Australia*

[2]*Centre for Research in Complex Systems,*

*Charles Sturt University, Albury, New South Wales, Australia*

[3]*Centre for Informatics and Applied Optimisation,*

*School of Science, Information Technology and Engineering, University of Ballarat, Victoria, Australia*

frada.burstein@monash.edu

*Abstract*— **Clinical decision-support is gaining widespread attention as medical institutions and governing bodies turn towards utilising better information management for effective and efficient healthcare delivery for quality assured outcomes. Amass of data across all stages, from disease diagnosis to palliative care, is further indication of the opportunities and challenges created for effective data management, analysis, prediction and optimization techniques as parts of knowledge management in clinical environments. A Data-driven Decision Guidance Management System (DD-DGMS) architecture can encompass solutions into a single closed-loop integrated platform to empower clinical scientists to seamlessly explore a multivariate data space in search of novel patterns and correlations to inform their research and practice. The paper describes the components of such an architecture, which includes a robust data warehouse as an infrastructure for comprehensive clinical knowledge management. The proposed DD-DGMS architecture incorporates the dynamic dimensional data model as its elemental core. Given the heterogeneous nature of clinical contexts and corresponding data, the dimensional data model presents itself as an adaptive model that facilitates knowledge discovery, distribution and application, which is essential for clinical decision support. The paper reports on a trial of the DD-DGMS system prototype conducted on diabetes screening data which further establishes the relevance of the proposed architecture to a clinical context.**

## I. INTRODUCTION

The challenge of accumulating, managing and exploring vast collections of heterogeneous data is commonplace in large medical organisations. Financial savings, quality assurance, workflow optimisation, error prevention, facilitation of best practices and medical research are some of the major expected benefits from proper clinical decision-support (CDS). As reported in [1], technological innovations have gradually expanded the boundaries of CDS and thereby its definition has been constantly changing in each decade [2]. CDS has evolved from medical data processing tools to complex decision support frameworks and infrastructures for clinical knowledge management. Two prominent taxonomies for CDS architectures are presented in [1] and [2]. In [2], the authors define architecture as the form of interaction between related systems, with four distinct CDS phases; standalone, integrated, standards-based and the service model. The phases are evolutionary but it is not uncommon to find systems from the first phase in contemporary clinical environments. Standalone systems would take clinical parameters as input and make suggestions of diagnoses or therapy; they were confined to a single specific area of medicine. In the integration phase CDS were incorporated into clinical information systems (CIS) resulting in proactive systems with more streamlined data entry. The requirements of inter-organisational and government policy-driven information exchange led to the next phase, the use of standards to represent, encode, store and share clinical knowledge. The more recent phase, the service model, separates CIS from CDS and provides interfaces for communication between the two. This allows a broad range of information interchange possibilities across departments and even hospitals, as well as other organisations concern. Despite obvious challenges and the need for agreed standards for information exchange, such model also provides wide opportunities for further clinical knowledge creation through mining data from various sources while recognising multiple perspectives of the users.

CDS architectures are explored in the context of the underlying technology [1]. The authors highlighted the need for an overarching architecture that integrates three existing architectures. These three, which are identified in terms of the underlying technology, are information management, data analytics and knowledge management. Convergence into a single overarching architecture is further justified as it can then address many of the grand challenges of CDS stated in [3]. Despite the emphasis on architecture and the need for convergence, existing literature does not present an integrated architecture that maximises throughput of accumulated data for better informed CDS.

Decision Guidance is a recent concept that aims to direct decision makers towards rational outcomes in complex environments exploiting the richness of dynamical accumulation of data. A Decision Guidance Management System (DGMS) was defined as a productivity platform for closed-loop data acquisition, learning, prediction and decision optimization [4]. The authors reported positive outcomes on

the application of DGMS to supply chain management, energy distribution and management [4-5].

In this paper we propose an extension to the DGMS architecture, the Data-Driven DGMS (DD-DGMS) that has the potential to address the convergence requirement identified in CDS. The proposed extension introduces a data warehouse as the intermediary layer between conventional data stores and decision support techniques. Although the data warehouse concept is widely used in business, it found less relevance to medical and biomedical informatics research and practice. We analyse the studies towards Clinical Data Warehouse (CDW) with the aim of demonstrating how advanced multi-dimensional modelling functionality facilitated by underlying data warehouse can help clinical scientists in dynamic hypothesis generation and testing. As elucidated in the rest of the paper, the plasticity of a data warehouse is instrumental in enabling multivariate decision guidance in clinical contexts, in particular in the context of translational research which aims to generate and test hypothesis based on collected data.

The rest of the paper is organised as follows. Section II presents the motivation for data-driven decision guidance from a clinical perspective with emphasis on the information needs of clinical scientists who are actively involved in all phases of the decision making process from initiation to conclusion. Section III delineates data warehousing, its underlying dimensional model and business intelligence (BI) techniques used to navigate and explore the content in a warehouse. This section also discusses the CDW systems reported in the literature. The proposed architecture is presented in Section IV. It outlines the features of the enhanced DGMS along with benefits in a clinical context. Section V reports results from a prototypical trial conducted on Diabetes screening data while Section VI sees to the conclusion of the paper.

## II. Motivation

Clinical scientists are responsible for elucidating factors associated with disease risk, identification and progression and providing a platform to medical professionals for the diagnosis and management of disease. They collaborate with medical practitioners to conduct research on disease management, largely based on data accumulated from clinical examinations. The declining role of clinical scientists in medical research has been identified as a potential reason for the critical gap (termed the 'valley of death' crisis) that lies between bench research and bedside treatment [6]. In [7], the authors emphasise the increasingly important role of clinical scientists in translational research; research that converts laboratory discoveries into clinical interventions.

Data analysis, interpretation and utilisation in practice is mostly restricted to the risk assessment based on multivariate regression modelling where the researcher or clinician decides *a priori* on features to be analysed and controlled for. A two group model such as diabetes versus non-diabetes explores the differences in features measured during clinical trials or research projects. This restricts the power of the information available in terms of identification of novel clinical/pathological interactions. In contrast, routine clinical data collection provides a much larger database and greater scope for investigating and understanding disease processes, how patients are currently treated and treatment outcomes. Motivation for the proposed framework stems from this need to explore, analyse and aggregate large datasets.

Further impetus toward a DD-DGMS approach comes from [8] which illustrates that data mining techniques with medical datasets lead to insights that often engendered a deeper understanding of conditions and informed the design of controlled medical experiments. However, although many statistical and data mining algorithms exist, few are designed to assist an analyst to explore large and complex datasets for the kind of insight into underlying knowledge that is required in a DD-DGMS. For instance, [9] revealed that by presenting knowledge in a form that medical specialists could find intuitively easy to assimilate, unexpected interactions were found to be extremely interesting and stimulated the creation of explanatory hypotheses. For example, that approach identified the absence of reflex in the knees and ankles together with a mid-range glucose reading was unexpectedly highly predictive of diabetes. This prompted analysts to hypothesise about possible nervous system dysfunction being present at a pre-diabetes stage. Medical experts suggested explanatory hypotheses; perhaps post-menopause hormonal regulation may be more influential in pre-diabetes than currently thought. Reflex and glucose level tests are easy to perform clinically, so this insight may lead to new ways to assess the risk of diabetes and diabetes complications from a preclinical stage.

A data-driven DGMS relies heavily on the storage and codification of data in a form that can readily lead to the insights and guidance desired. This form should also provide means of easily combining features and investigating complex connectivity patterns between features associated with disease progression. A data warehouse for this purpose is introduced in the next section.

## III. Data Warehousing Basics

Kimball broadly defines a data warehouse as a copy of transaction data specifically structured and optimized for queries and analysis [10]. Despite the introduction of data warehouses for clinical data more than two decades ago, very few institutions have actually built warehouses [11]. Challenges include the difficulties inherent in developing a meta-model that encompasses a broad range of variables and values structured along a number of dimensions.

The dimensional model is the conceptual basis for data warehouse development. Dimensional model design is essentially a denormalization technique that provides an intuitive view of data corresponding to the main areas of interest of the problem space [12]. It is commonly represented as a subject-oriented structure composed of fact tables and dimensional tables. The fact table contains keys and numerical measures that can be summarized in terms of dimensions.

Dimensions are typically composed of attributes and hierarchies that identify areas of interest in the problem domain. The fact table is linked to all dimensional tables resembling a star or snowflake structure (Fig 1).

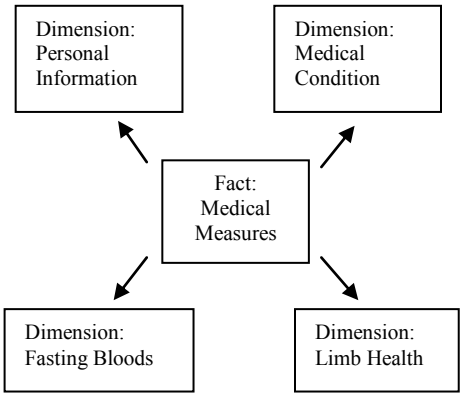


Fig. 1 A dimensional model for a Clinical Data Warehouse

The data warehouse organizes attributes of each input vector into the star or snowflake schema of the dimensional model.

In the past decade, data warehousing has emerged as a widely used platform for business intelligence (BI) capabilities that enable strategic reporting and decision making. Despite its prevalence in many industries, its adoption by medical organizations has been limited. Early implementations of CDW were aimed at addressing specific clinical problems. There are still a few recent case studies that demonstrate the applicability of data warehouse concept in medical domain. For example, in [13], the authors describe the use a data warehouse for hospital infection control. It was populated with data from three hospitals and demonstrated to be useful for measurement of antimicrobial resistance, antimicrobial use, the cost of infections, and detection of antimicrobial prescribing errors. In [14], the authors present a review of the Enterprise Data Trust at the Mayo Clinic, which is a collection of all electronic data organized to support information management, analytics and high-level decision-making.

In recent research endeavours [15-16], the authors have proposed and implemented data warehousing solutions to address the information needs of translational research. In [15], the authors developed a data warehouse integrating pathologic and molecular data with a clinical data model to support a breast cancer translational research program. STRIDE (Stanford Translational Research Integrated Database Environment) database environment [16], is an informatics platform for clinical and translational research. It consists of a data management system, a CDW and a development framework for new applications.

The wide acceptance and implementation of CDW systems is a strong indicator of its usability in clinical contexts. The following section illustrates the novelty of the proposed architecture, where a data warehouse is introduced into DGMS as an intermediary layer that facilitates multivariate decision guidance.

## IV. Architecture and Methodology

The DGMS architecture [4] was designed to be used in iterative loop-back phases. The first phase uses the database and domain knowledge to define a data space from which knowledge is derived (learned). In the second phase learning and domain knowledge are used for prediction and simulation. Prediction and simulation outcomes are used for decision optimization in the third phase, while in the final phase data acquisition queries are used as feedback to reduce ambiguity of decisions. The authors introduce DG-SQL (an extension of SQL) as a query language to support and enable the phases of operation in DGMS.

The proposed DD-DGMS architecture replaces DG-SQL based intermediation with a data warehouse, whereby existing features of the framework are enhanced and further functionality is introduced. In order to discuss the functionality we distinguish between two groups of users found in a clinical environment based on their information needs. The first group comprises of users (operational level) interested in short term outcomes such as doctors investigating medication usage, clinical scientists seeking better means to reach diagnoses or determine effectiveness of medication, physical activity and diet in risk reduction of disease progression. The second group of users (strategic level) such as clinical administrators and policy makers seek information relevant for optimising treatment regimen that have the best individual outcomes by reducing disease progression and disease associated mortality within the economic constraints of the current health care system.

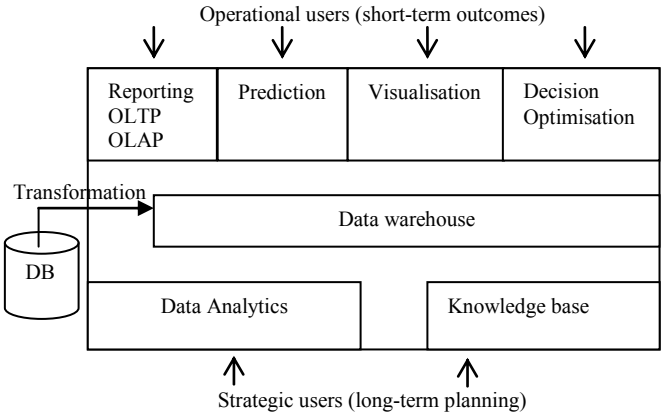


Fig. 2 Architecture of the Data-driven Decision Guidance Framework

The DD-DGMS architecture, its features and its interactions with the two groups of users are illustrated in Fig 2. The use of each feature is not strictly limited to a single group, for instance strategic users would benefit from decision optimisation as equally as operational users. As noted earlier, the data warehouse is the intermediary layer facilitating each feature and operational phase. Its flexible structure enables multivariate information retrieval and incorporates user

feedback for subsequent use. Such interaction between the operational and strategic levels creates an opportunity for acquiring new and refining existing knowledge as part of the continuous knowledge management cycle. The operation of the architecture in a typical clinical setting is outlined in the following discussion of each of its components.

*Data Transformation*

A variety of electronic data stores can be found in a typical clinical environment. Flat file storage, multiple database vendors and different data models are common issues in any organisation where data from different sources, sub-units or departments needs to be integrated into a single structure. Many of the common technical challenges to data integration are discussed in detail elsewhere [10]. We focus on three issues specific to clinical data integration; they are discretisation, temporal abstraction and cardinality.

*1) Discretisation:* Numeric clinical measures are continuous (real) by nature. For aggregation and analysis that supports decision guidance these measures need to be converted to discrete values or ranges. In most circumstances the discretisation criteria is determined by clinicians. Where domain expertise is unavailable an algorithmic approach is adopted. A typical discretization process attempts to seek a compromise between information quality and statistical quality and broadly consists of four steps: sorting the continuous values, evaluating a cut-point for splitting or merging, splitting or merging intervals of the continuous value and termination [17].

*2) Temporal Abstraction:* Patient monitoring data consists of time-stamped variables. The purpose of temporal abstraction is to derive high-level qualitative descriptions from such low-level quantitative measures of a variable [18]. Given the multivariate nature of clinical data spaces, it is important to ensure temporal abstractions do not conflict with each other. The availability of qualitative descriptions improves the decision guidance process as they are context sensitive and relevant to actionable outcomes.

*3) Cardinality:* Cardinality is temporal abstraction applied to a group of variables that have a contextual association. A prominent example for the use or cardinality, which is also illustrated in the experiments section of this paper, is diagnosis of the time course of a medical condition. A patient with a chronic disease would have frequent tests conducted on variables indicative of the stage of the disease, the actual measurements are candidates for temporal abstraction while cardinality is used to identify each individual test.

Once the data is integrated from multiple sources and transformed into a clinical-context sensitive format, it is uploaded into the warehouse.

*Data Warehouse*

The development of a data warehouse starts with the dimensional design process. As stated in Section II, numerical measures in the problem domain are identified initially. The granularity of these can vary from coarse-grained (e.g. number of patients) to fine-grained (level of blood glucose). The dimensions of interest are identified next. As shown in Fig 1 these can be any aspect of the clinical environment that the numeric measures need to be analysed or aggregated from, such as personal details of the patient, exercise routine medical conditions and test results. It is necessary to introduce cardinality as a dimension when temporal dimensions such as medical conditions and test results (that change over time for the same patient) are included. Further dimensions are introduced to capture user feedback. Information on aggregates and trends derived by clinicians as well as clinical outcomes can be translated back to the warehouse as dimensions to be used in future analysis.

*Reporting- OLTP and OLAP*

Reporting is essentially the execution of queries on a structured data store. The underlying dimensional model of the warehouse supports both Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP). The latter is more relevant for analysis and aggregation where data cubes can be formed by introducing multiple dimensions to the query. Furthermore, slicing and dicing operations can be performed on a cube to increase/decrease granularity of a multivariate query. Multidimensional expressions (MDX), the query language for OLAP can also be used for reporting.

*Prediction*

The availability of time-course analysis capabilities allows a clinician to use the warehouse to predict the subsequent phase of a patient affected by a medical condition based on past records of other patients in similar circumstances. Even well known disease trajectories can be validated with the DD-DGMS approach.

*Visualisation*

While OLTP and OLAP are successful at aggregation and analysis, the large number of dimensions in clinical settings can require visualisation features for improved understanding. Groups of patients at the edges of overlapping dimensions are easily identified visually than by any other means.

*Decision Optimisation*

Decision optimization is partially the validation of the outcomes obtained from prediction and reporting features. Given the dimensions in a warehouse are independent to each other, outcomes can be reviewed by removing existing or adding further dimensions. Optimal aggregates would be consistent regardless of the changes to dimensions.

*Data Analytics*

Data analytics is the first feature that's more applicable for long-term planning. Cubes of data that are of interest to the clinical scientist can be isolated using OLAP and further analysed using data mining algorithms. There are a variety of data mining algorithms to address different requirements such as classification, association and clustering. Trends and patterns prevalent in a subset of the population are valuable findings that can guide long term decision-making and policy shifts. Without the support of an underlying warehouse, it would be very difficult to identify and isolate such subsets.

*Knowledge Base*

Outcomes from all the above features are the building blocks of knowledge and understanding that can be derived from accumulated data. Use of the proposed DD-DGMS architecture by clinical scientists would thus generate valuable outcomes that can support clinical practice as well as research. These outcomes are initially maintained within the warehouse and transferred into a knowledge base when sufficient data-based evidence is accumulated. A mature knowledge base can be useful to address knowledge management concerns such as ontology generation, training and guidelines development.

## V. Experiments and Results

A prototypical trial was conducted to establish the feasibility of the proposed DD-DGMS architecture in biomedical context. The dataset derives from the Diabetes Screening Complications Research Initiative (DiScRi) conducted at a regional Australian university [19]. It is a diabetes complications screening program in Australia where members of the general public participate in a comprehensive health review. The screening clinic has been collecting data over ten years and includes over one hundred features including demographics, socio-economic variables, education background, clinical variables such as blood pressure, body-mass-index (BMI), kidney function, sensori-motor function as well as blood glucose levels, cholesterol profile, pro-inflammatory markers, oxidative stress markers and use of medication. Data on 273 attributes from over 2500 attendances of nearly 900 patients have been collected in recent years. The dataset has been used in several data mining applications [9, 20-21]. Application of the DD-DGMS approach to this dataset is discussed in the following subsections.

### A. DiScRi Data Transformation

Data transformation initiated with the replacement of missing values, erroneous values and records. Given the clinical nature of the dataset, the three issues of discretisation, temporal abstraction and cardinality were also addressed in this phase. Discretisation was applicable to a large number of the attributes. In most cases the clinical scientist was able to provide a clinical discretisation scheme (Table 1).

TABLE I
Examples of Clinical Discretisation Schemes

| *Attribute* | *Description* | *Clinical discretisation scheme* |
|---|---|---|
| Age | Participant's age on test date | <40, 40-60, 60-80, >80 |
| Diagnostic HT Years | Number of years since diagnosis of hypertension | <2, 2-5,5-10, 10-20, >20 |
| FBG | Fasting blood glucose level | <5.5 very good, 5.5-6.1 high, 6.1 to 7 preDiabetic, >=7 Diabetic |
| Lying DBP Average | Diastolic blood pressure when lying down | < 60 = low, 60-80 = normal, 80-90 = high normal, >90 = hypertension |

Attributes without clinical schemes were duplicated with one having the original continuous form and the other discretised using a top-down or bottom-up discretization technique [17] that was most relevant to the type of data. Temporal abstraction and cardinality was required by many attributes in the dataset and thereby warranted a separate dimension in the dimensional model. This is discussed in following subsection.

### B. DiScRi Data Warehouse

Domain expertise was necessary to determine the key dimensions of interest from the total of 273 attributes (Fig. 3). As individual patients attended the screening clinic multiple times, it was necessary to introduce cardinality as a separate dimension. Except for Personal Information (which contained attributes such as gender, family history of medical conditions), all other dimensions were composed of attributes recorded for each visit/test. While the fact table would distinguish between records, the cardinality dimension was necessary to distinguish between patients.

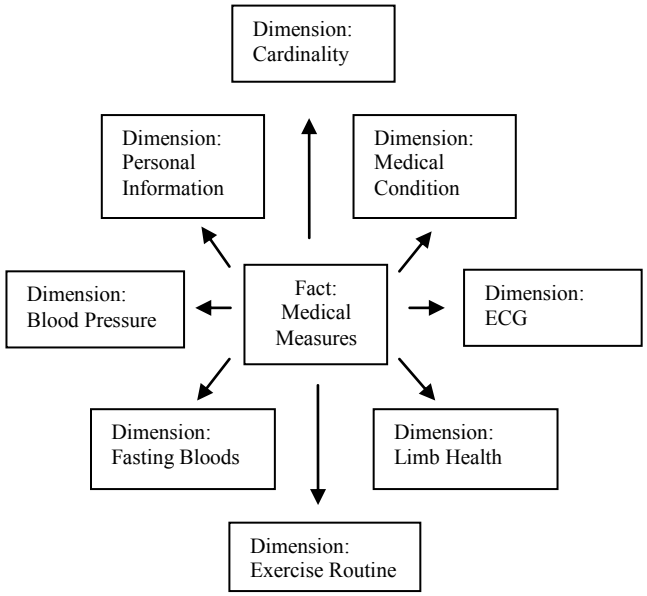


Fig. 3 Dimensional model used in the prototypical trial

The proposed dimensional model (Fig. 3) was used to construct the data warehouse. Microsoft SQL Server [22] was selected as the database platform for its ease of use for rapid development. Microsoft Business Intelligence Studio [22] was used to populate the warehouse with the pre-processed DiScRi dataset and also as an analysis front end.

### C. DiScRi OLAP Reporting

Upon populating the data warehouse, it was possible to execute OLAP queries using both the graphical user interface and MDX. The graphical interface is shown in Fig. 4. The ease of use of this interface by non-technical users (such as clinical scientists) is noteworthy. The measures and attributes (grouped by dimensions) are listed on the left panel.

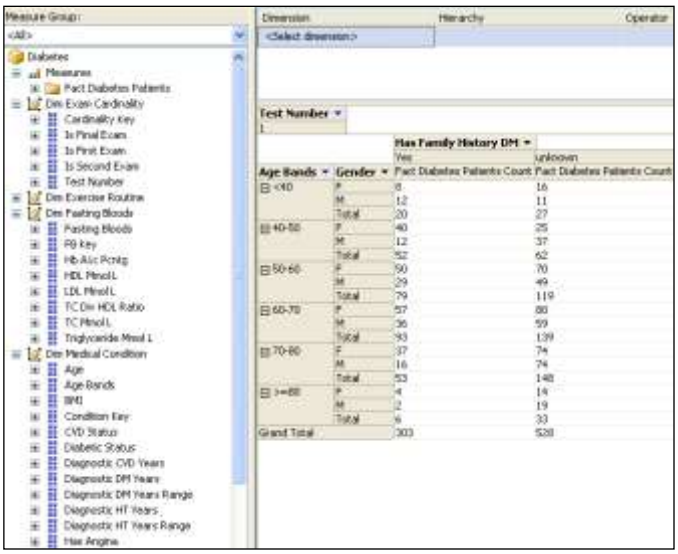

Fig. 4 'Drag and Drop' features for query construction in Microsoft BI Studio

They can be dragged into the centre query area to dynamically generate queries and view the aggregated results. Fig. 4 shows the family history of diabetes by age group and by gender. Any number of attributes can be introduced to the query area for drill-down and roll-up features.

A graphical outcome is shown in Fig.5, where an OLAP query plots the age and gender distribution of patients with diabetes.

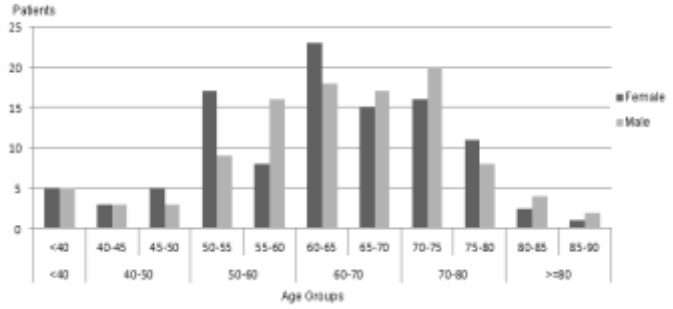


Fig. 5 OLAP outcome: Age and gender (females – left column) distribution of patients with diabetes

Using the drill-down feature, age distribution is shown at two levels of granularity. This has exposed a distinction between genders in the 70-80 age group; males dominate the 70-75 subgroup while females are the majority in the 75-80 subgroup.

Also of note in Fig 5 is the reduction in proportion of females with diabetes in the older age groups. Hesslera et al identified patient age as a neglected factor in diabetes management [23]. They illustrated that younger age groups (21-45 and 46-65) were associated with lower diabetic self-efficacy and higher stress compared with patients in the 65-80 age group. This led them to recommend education programs that targeted younger groups. Although gender was not found to be related to age in that study, the OLAP outcome of Fig 5 suggests that there may well be a gender effect within the older age group as the proportion of women with diabetes drops substantially over 78. This result provides an illustrative example of the potential the DD-DGMS approach advanced here has in enhancing translational research.

Another issue in translational research surrounding diabetes screening and management includes the identification of improved methods for performing a risk assessment of cardiovascular autonomic neuropathy, a well-known complication of diabetes. Tests known as the Ewing battery of tests have been advanced as simple clinical procedures that lead to a risk assessment [24]. However, some of the procedures such as the hand grip test cannot be applied to the elderly because of arthritis or other reductions in capacity. A DD-DGMS approach enables the data to be accessible to drive decision guidance hypothesis formulation regarding other patient characteristics that could be used in place of the missing test. Results of this analysis are currently being explored.

Fig. 6 illustrates the use of a clinical discretisation scheme (DiagnosticHTYears) in an OLAP query.

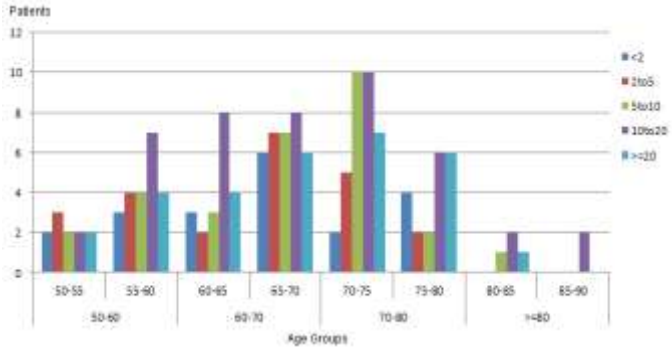


Fig. 6 OLAP outcome: Distribution of number of years since diagnosis of hypertension by age groups.

Patients with hypertension are identified by their age groups and by the number of years since diagnosis of hypertension. Again, the use of drill-down feature in age groups detects a significant drop in the number of 5-10 year hypertension cases in the age sub-groups of 70-75 and 75-80.

## VI. Discussion and Conclusion

DGMS and data warehousing provides the means for seamless integration and dynamic combination of features for investigating complex connectivity patterns between features associated with disease progression. To enhance translational research, data obtained from the annual screening clinic needs to be interpreted in terms of outcome measures following diverse treatment options present in the screening cohort. New associations between personal health status, intervention and individual outcome that also reflects a wider population use is an essential part of current health care research.

The paper reports on a trial of the DD-DGMS prototype conducted on diabetes screening data which further establishes the relevance of the proposed architecture to a clinical context. This paper only reports on outcomes from operational use of the DD-DGMS architecture. The strategic benefits are generated by long-term usage of the architecture

and are of significant value in terms of knowledge management in a clinical setting. This is part of the current and future work the authors are involved in. It is envisaged that access to the infrastructure for decision guidance and clinical decision support facilitated by multi-dimensional modelling will equip the clinical scientists to produce more refined and better informed test plans for future data collection and analysis.

We argue that such an approach can make a significant contribution to creation, use and re-use of medical knowledge thus facilitating effective clinical practice and better support for medical decision-making.

## ACKNOWLEDGMENT

This research is partly supported by the University of Ballarat 'Self-sustaining Regions Research and Innovation Initiative', an Australian Government Collaborative Research Network (CRN).

## REFERENCES

[1] A. Sen, A. Banerjee, A. P. Sinha, and M. Bansal, "Clinical decision support: Converging toward an integrated architecture," Journal of Biomedical Informatics, vol. 45, no. 5, pp. 1009–1017, Oct. 2012.

[2] A. Wright and D. F. Sittig, "A four-phase model of the evolution of clinical decision support architectures," International Journal of Medical Informatics, vol. 77, no. 10, pp. 641–649, Oct. 2008.

[3] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates, "Grand challenges in clinical decision support," Journal of Biomedical Informatics, vol. 41, no. 2, pp. 387–392, Apr. 2008.

[4] A. Brodsky and X. S. Wang, "Decision-Guidance Management Systems (DGMS): Seamless Integration of Data Acquisition, Learning, Prediction and Optimization," Hawaii International Conference on System Sciences 2008, pp. 71–71.

[5] J. Arinez, S. Biller, A. Brodsky, D. Menasce, Guodong Shao, and J. P. Sousa, "Decision-Guided Self-Architecting Framework for integrated distribution and Energy Management," 2011, pp. 1–8.

[6] D. Butler, "Translational research: crossing the valley of death," Nature, vol. 453, no. 7197, pp. 840–842, Jun. 2008.

[7] S. F. Roberts, M. A. Fischhoff, S. A. Sakowski, and E. L. Feldman, "Perspective: Transforming science into medicine: how clinician-scientists can build bridges across research's 'valley of death'," Academic Medicine, vol. 87, no. 3, pp. 266–270, Mar. 2012.

[8] R. Rosales, and R.R. Bharat, "Guest Editorial: Special Issue: Impacting Patient Care by Mining Medical Data", Data Mining and Knowledge Discovery. Volume 20. No.3, 2010

[9] A. Quinn, A. Stranieri, J. Yearwood, G. Hafen, and H.F. Jelinek, "AWSum-Combining Classification with Knowledge Acquisition". International Journal of Software and Informatics Vol 2(2), 2008, pp 199-214.

[10] R. Kimball, "The Data warehouse lifecycle toolkit". Indianapolis: Wiley publishing, 2008.

[11] M. de Mul, P. Alons, P. van der Velde, I. Konings, J. Bakker and J. Hazelzet. "Development of a clinical data warehouse from an intensive care clinical information system" Computer Methods and Programs in Biomedicine, Volume 105, Issue 1, January 2012, Pages 22-30.

[12] R. Agrawal, A. Gupta, and S. Sarawagi, "Modeling multidimensional databases," in Proceedings of the 13th International Conference on Data Engineering, 1997, pp. 232–243.

[13] M. F. Wisniewski, "Development of a Clinical Data Warehouse for Hospital Infection Control," Journal of the American Medical Informatics Association, vol. 10, no. 5, pp. 454–462, Jun. 2003.

[14] C. G. Chute, S. A. Beck, T. B. Fisk, and D. N. Mohr, "The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data," Journal of the American Medical Informatics Association, vol. 17, no. 2, pp. 131–135, Feb. 2010.

[15] H. Hu, M. Correll, L. Kvecher, M. Osmond, J. Clark, A. Bekhash, G. Schwab, D. Gao, J. Gao, V. Kubatin, C. D. Shriver, J. A. Hooke, L. G. Maxwell, A. J. Kovatich, J. G. Sheldon, M. N. Liebman, and R. J. Mural, "DW4TR: A Data Warehouse for Translational Research," Journal of Biomedical Informatics, vol. 44, no. 6, pp. 1004–1019, Dec. 2011.

[16] The Stanford translational research integrated database environment(STRIDE). <http://clinicalinformatics.stanford.edu/STRIDE/>.

[17] S. Kotsiantis and D. Kanellopoulos, "Discretization Techniques: A recent survey", GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 47-58

[18] M. Stacey and C. McGregor, "Temporal abstraction in intelligent clinical data analysis: A survey," *Artificial Intelligence in Medicine*, vol. 39, no. 1, pp. 1–24, Jan. 2007.

[19] H.F. Jelinek, C. Wilding, and P. Tinley, "An innovative multi-disciplinary diabetes complications screening programme in a rural community: A description and preliminary results of the screening", Australian Journal of Primary Health 12, 2006, pp 14–20.

[20] D. Cornforth, and H. Jelinek, "Automated classification reveals morphological factors associated with dementia". Applied Soft Computing 8, 2007, pp 182–190.

[21] S. Huda, H. Jelinek, B. Ray, A. Stranieri, and J. Yearwood, "Exploring novel features and decision rules to identify cardiovascular autonomic neuropathy using a Hybrid of Wrapper-Filter based feature selection", International Conference on Intelligent Sensors, Sensor Networks and Information Processing ISSNIP 2010 pp297-302.

[22] Microsoft SQL Server BI Toolkit http://www.microsoft.com/sqlserver/en/us/solutions-technologies/SQL-Server-2012-business-intelligence.aspx

[23] Danielle M. Hesslera, Lawrence Fishera, Joseph T. Mullan, Russell E. Glasgow, Umesh Masharanid. 2011. Patient age: A neglected factor when considering disease management in adults with type 2 diabetes. Patient Education and Counseling Volume 85, Issue 2, , Pages 154–159

[24] D. Ewing, J. Campbell, B. Clarke, 1980. The natural history of diabetic autonomic 425 neuropathy, Q. J. Med. 49 (1980) 95–100