8-1-2000

Cancer Surveillance using Data Warehousing, Data Mining, and Decision Support Systems

Guisseppi A. Forgionne
University of Maryland - Baltimore County

Aryya Gangopadhyay
University of Maryland - Baltimore County

Monica Adya
Marquette University, monica.adya@marquette.edu

Published version. *Topics in Health Information Management*, Vol. 21, No. 1 (August 2000): 21-34.

Topics in Health Information Management
Issue: Volume 21(1), August 2000, pp 21-34

Publication Type: [Contemporary Issues in Health Information Management]
ISSN: 1065-0989
Accession: 00008479-200008000-00003
Keywords: breast cancer surveillance, data mining, data warehousing, decision support systems

[Contemporary Issues in Health Information Management] Previous Article | Table of Contents | Next Article

Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems

Forgionne, Guisseppi A. PhD, MBA, MA, BS; Gangopadhyay, Aryya PhD, MS, B.Tech; Adya, Monica PhD, MS, BS

Author Information

Professor of Information Systems; Information Systems Department; University of Maryland Baltimore County; Catonsville, Maryland (Forgionne)

Assistant Professor; Information Systems Department; University of Maryland Baltimore County; Catonsville, Maryland (Gangopadhyay)

Assistant Professor; Information Systems Department; University of Maryland Baltimore County; Catonsville, Maryland (Adya)

Back to Top

Abstract

This article discusses how data warehousing, data mining, and decision support systems can reduce the national cancer burden or the oral complications of cancer therapies, especially as related to oral and pharyngeal cancers. An information system is presented that will deliver the necessary information technology to clinical, administrative, and policy researchers and analysts in an effective and efficient manner. The system will deliver the technology and knowledge that users need to readily: (1) organize relevant claims data, (2) detect cancer patterns in general and special populations, (3) formulate models that explain the patterns, and (4) evaluate the efficacy of specified treatments and interventions with the formulations. Such a system can be developed through a proven adaptive design strategy, and the implemented system can be tested on State of Maryland Medicaid data (which includes women, minorities, and children).

Back to Top

INTRODUCTION

A long-term public objective is to reduce the incidence of cancer and the morbidity, mortality, and costs associated with the national cancer burden or the oral complications of cancer therapies, especially as related to oral and pharyngeal cancers. 1 To achieve this objective, it first will be necessary to monitor populations; collect relevant cancer screening, incidence, treatment, and outcomes data; identify pertinent cancer patterns; explain the patterns; and

Outline

translate the explanations into effective diagnoses and treatments.

Data for assessing the diverse aspects of cancer surveillance are being captured in divergent formats and stored throughout many organizations. Sources of these data include cancer registries; fee-for-service insurance bills; Medicare, Medicaid, and private managed care encounter reports; multiple payer databases; and discharge summaries.

The various formats and locations create burdensome tasks for clinicians, administrators, policy makers, researchers, and analysts seeking to assess the national cancer burden from the available data. Interested parties must identify the data pertinent to their analyses and evaluations. Relevant data must be located, extracted, and captured. These captured data must be converted into the variables desired by the interested parties. Patterns in the data must be identified and discerned. Explanatory models must be suggested for the patterns and the data converted into the variables relevant to the explanatory models. The models must be tested against the collected and processed cancer surveillance data.

Such data collection, processing, and analysis are time consuming and costly. Success is highly dependent on the abilities, skills, and domain knowledge of the interested parties. Even the most talented and skilled parties have incomplete knowledge about the study domain, pertinent information technology (IT), and relevant analytical tools. Data sharing across interested groups is limited. Consequently, much useful information may be lost in the cancer surveillance effort.

IT and analytical approaches are available to assist interested parties in collecting, processing, and analyzing cancer data. There are methods to view the various secondary sources, extract the relevant data components, capture the elements, and warehouse the captured data. 2,3 Data mining techniques exist to access the data warehouse and detect care, outcome, and therapy patterns. 4,5 There are statistical methodologies to develop models that explain the detected patterns. 6,7 Decision support systems are available to readily deliver the methods, techniques, methodologies, and developed models to the interested parties. Providing such warehousing, data mining, and decision support can enhance cancer surveillance efficiency and effectiveness significantly.

This article develops a framework that demonstrates how decision support can be achieved through an integration of the relevant methodologies and technologies. There is a discussion of cancer surveillance through IT. Next, an information system is presented that can assist practitioners and policy analysts effectively and efficiently assess the cancer burden from health claims data. The article concludes with a study of the implications for health care research and practice.

Back to Top

CANCER SURVEILLANCE THROUGH IT

The cancer surveillance process with IT is summarized in Figure 1. Pertinent cancer data is warehoused. Statistical and artificial intelligence data mining techniques are used to detect cancer patterns in the warehoused data. Additional statistical methodologies are used to: (1) identify variables correlating with the patterns, (2) formulate hypotheses of variables that cause cancer, and (3) test the hypotheses. To some extent, these evaluations are guided by, and rely on, the judgment, insights, and experience of the researcher or analyst. An information system delivers the data, models, analyses and evaluations, and created knowledge needed for decision support.

Figure 1

Back to Top

IT


Figure 2

The information system will consist of the components shown in Figure 2. The system, labeled as the Cancer Surveillance System (CSS), will consist of a geographical information system (GIS), an executive information system (EIS), and a decision support system (DSS). Besides extracting data and creating thematic maps, the prototype GIS provides inputs for the EIS. These inputs consist of the geographic, health outcome, environmental and natural resource, demographic, and health care variables associated with the desired geographic area.

The EIS provides a database management system (DBMS) and an intelligent decision support system processor (IHP). The DBMS extracts geographic area conditions from the GIS, takes ad hoc queries from the user in an interactive manner, displays the query results in attractive reports, and stores the information in a data warehouse. The IHP captures the DBMS-generated data, updates the DSS's spatial and temporal statistical models, performs DSS analyses and evaluations, and generates detailed reports of the results automatically, without human (manual) intervention. The suggested approach is to obtain the data from the various sources, create a data warehouse, and have an EIS perform access and reporting from the warehouse.

Back to Top

Data warehousing and mining


Figure 3

The system will provide an open and scalable online analytical processing (OLAP) solution using the World Wide Web with the thin-client architecture shown in Figure 3. The clients generate and view reports and graphs of data stored in the multidimensional database server, which derives the raw data from the data warehouse at the database server. Users can run queries and generate reports using any HTML-based browser without having an application session run on the client. This architecture reduces the cost of client-server management by centralizing the data and applications on the server, and the architecture makes it simple to distribute and upgrade applications and add new OLAP users.

Data mining techniques will then be used to extract cancer patterns in the data. Such techniques will rely heavily on artificial intelligence technologies, in particular neural networks and genetic algorithms, 8 and statistical methodologies. This mining can facilitate information analysis using either predictive or descriptive modeling. 9 Descriptive modeling is exploratory in nature and contributes to the discovery of previously unknown patterns, trends, and associations in the data. Predictive modeling allows the examination of data in a more traditional way by testing specific hypotheses. For instance, health care providers may anticipate an increased incidence of breast cancer in areas that have radioactive waste disposal units. By relying on historical data, data mining techniques can test for this hypothesis in areas where such waste disposal is prevalent.

The data mining is expected to reveal important patterns in several areas, such as factors affecting cancer care, cancer recurrence, and cancer prevention. These patterns are expected to be revealed at the population, subpopulation, and individual levels. Explanatory and other statistical models will then be used to refine patterns that emerge from the use of these techniques. Further, such validated findings will eventually allow users to develop rules that can be encoded into the CSS.

Back to Top

Conceptual CSS architecture


Figure 4

The CSS will have the conceptual architecture shown in Figure 4. As the figure illustrates, interested parties utilize the system's computer technology to interactively process inputs into desired outputs.

Back to Top

Inputs

The CSS has a database that captures and stores geographic, demographic, environmental, health outcome, and health care data. A model-base (organized repository of models and algorithms) captures and stores GIS, health, and statistical models required and desired by interested parties. These models compute and estimate cancer-relevant variables; describe and explain the spatial, temporal, and space-time relationships between the variables and cancer development; and use the relationships to simulate overall and oral cancer incidence and mortality in general and special populations. In addition, a knowledgebase captures and stores spatial and temporal profiles linked to cancer incidence, development, and mortality patterns.

A multiple equation statistical model will be used to establish independent and joint causation between the cancer patterns and their determinants. This model would predict why the spatial patterns are occurring for specific demographic, spatial, and other profiles. The model would also determine if there are deviations from the profile, and therefore if there is an epidemic at hand, and identify the specific causes of the epidemic.

Back to Top

Processing

Interested investigators and officials would use the CSS interactively to perform analyses and evaluations, including the following:

* organizing data into parameters needed for the spatial, temporal, and space-time cancer analyses

* structuring models that represent and simulate overall and oral cancer development patterns in an integrated and complete manner

* simulating overall and oral cancer incidence and mortality under specified health care, health outcome, demographic, and environmental profiles

Profile data will be available from public and private research institutions.

The embedded GIS serves as a front end to the processing of the profile data. The system uses data mining to extract patterns of breast cancer occurrence and incidence. For instance, data mining techniques can identify certain regional characteristics that appear to have an impact on higher incidence of breast cancer in an area. The system links these findings and data to spatial dimensions and transfers the linked data to the embedded EIS.

An implanted EIS is used to (1) filter the profile data, (2) form CSS's database, (3) focus the filtered information, and (4) communicate deviations from expected cancer development patterns among affected parties. Such processing will involve the following:

* masking individual identities from publicly released health data

* disease surveillance

* identification of stabilized rates

* exposure assessment

* genetic activity profiling

* interactive spatial data analysis

* kriging

* scanning of statistical drill downs

* disease rate small area variation analyses

* space-time clustering

* standard mortality rate analyses

The EIS also will provide the focused data needed for the DSS analyses and evaluations. The DSS and Expert Systems guide the investigator through the intelligent modeling and the modelbase, database, and knowledgebase management needed for the GIS, EIS, and DSS analyses and evaluations.

Back to Top

Outputs

By controlling processing tasks in the desired way, the user can generate the following:

* visually attractive tabular and graphic claims status reports that describe the provider's clinical and administrative environment, track meaningful trends, and display important patterns

* claims condition forecasts

* provider policy and payer program simulation results

* recommended claims actions

The system also depicts, in a graphic manner, the reasoning (explanations and supporting knowledge) that leads to the suggested actions. By making desired CSS selections, the user will be able to simulate expected overall and oral cancer incidences and mortalities in general and special populations, compare actual results with the expected occurrences, and obtain a brief explanation for the deviations.

Users would enter the system and display a map depicting what is occurring in the area and what would be expected to occur in the absence of an epidemic. They would drill down to a very concentrated area to determine whether the deviations are focused in one area or are general to the entire region. Clustering models could be used manually to identify the pattern, or the user could have the system identify the pattern through artificial intelligence. Next, the user could request an explanation for the detected pattern. Such an explanation would call up the multiple equation statistical model, which would analyze whether the incident is caused by a specific factor or a series of interrelated factors specific to the region. With this information in hand, users could perform sensitivity analyses to find out what would happen if certain conditions change or if policies were changed. All results could be displayed on the maps with or without tabular reports, thereby justifying the analyses and evaluations.

Back to Top

Feedback Loops

Feedback from the processing provides additional data, knowledge, and enhanced decision models that may be useful for future cancer surveillance activities and tasks. Output feedback (often in the form of sensitivity analyses) is used to extend or modify the original analyses and evaluations. All processing

(including each feedback loop) is done in a user-friendly manner, with artificial intelligence (mainly expert system) technology, that meets the decision styles and requirements of participating parties.

Back to Top

EVALUATION PLAN

In concept, the CSS offers promise. This promise, however, must be evaluated. Measures must be developed to assess the effectiveness of the proposed system in reducing the national cancer burden. A research design and methodology must be developed and implemented to test CSS effectiveness against these measures. This section presents such an evaluation plan.

Back to Top

System effectiveness

Decisions will have to be made in the cancer surveillance process. Relevance must be defined for data extraction and pattern detection. Pertinent variables for the explanatory models must be identified. Tests of the models must be established, executed, and interpreted. The interpretations must be used to revise and update the models. Diagnostic and treatment outcomes must be constituted. A termination point for the analyses and evaluations must be fixed.


Figure 5

The CSS, then, must support the process of, as well as outcomes from, cancer surveillance decision making. Put another way, this surveillance involves a multiple-criteria assessment of effectiveness. Further, one general measure (process) is at least partially intangible in nature. For such situations, system effectiveness can be evaluated on the basis of the multiple criteria model shown in Figure 5. 10,11

As Figure 5 illustrates, the overall criterion, composite decision value, is determined by the process support offered by, and the outcome from using, the system. Outcome, in turn, is set by cancer surveillance performance and decision-maker maturity (increase in decision-making ability, skills, and domain knowledge), and process support is prescribed by phase and step proficiency, personal efficiency, and personal productivity. Several measures may fall within these general categories. The alternatives will be the list of provided support tools (none, CSS).

Equal or different weights can be assigned to each assessment variable. Variables not included in the analysis would receive weights of zero. Similarly, each support tool can receive an equal or disparate weight in the overall evaluation. In a completely impartial scheme, each variable would receive an equal weight, and each tool would have the same weight. By aggregating the estimated weights through the hierarchy, the evaluator will obtain an overall rating for each provided support tool. The overall criterion ratings then provide a basis for identifying decision value from the provided support tool. Highly effective tools will receive the largest decision values.

Back to Top

Research design and methods

The CSS will be developed through a proven adaptive design strategy. The research team will initially develop a working prototype of the system. Working in conjunction with the University of Maryland Baltimore County's (UMBC's) Center for Health Program Development and Management (CHPDM) health-knowledgeable personnel, the researchers next will review each evolving rollout of the system. Comments and suggestions for enhancements will be tested and evaluated in real time, typically in prototype form using rapid application development toolkits available within the development software. The researchers will make detailed

technical changes over a relatively short period of time (typically a few weeks), and the system changes will be presented to the officials for review and evaluation. The evolutionary process will continue until the latest version of the system is finalized.

Ideally, the research team would prefer to involve external interested parties in the adaptive design strategy. Such external parties are available from the University of Maryland Medical, Dental, and other health-related schools. However, these university personnel are very busy, and their involvement would divert their attention from their health duties. Moreover, time and budget limits make it more feasible to use UMBC CHPDM personnel. These CHPDM personnel have considerable health care experience and are familiar with the cancer surveillance issues. In addition, the specific aim in this project is to demonstrate the utility of the proposed approach, and the proposed adaptive design strategy will accomplish the specific aim. Even if utility is demonstrated, follow-up studies will be needed to confirm or refute this study's findings.


Table 1


Table 1


Table 2

The implemented system will be tested on State of Maryland Medicaid claims data. These data, which include medical, dental, rehabilitative, and other health care services claims by children, minorities, women, and the underserved, are currently managed for the state by UMBC's CHPDM. Nine years of these data, including various profiles categorized in the CHPDM data warehouses, are available. Table 1 gives an overview of the CHPDM files and Table 2 summarizes NIH-relevant profiles.

In the testing, the research team will scientifically evaluate CSS operations and results. Warehoused data will be compared with actual values. The research team will also identify available spatial and spatial-temporal variables, model components, complete models, and methodologies that, the literature suggests, predict cancer incidence and mortality. These simulation models will be captured and stored within the CSS. The system will automatically access the warehouse, extract literature-pertinent data, form the variables needed for the simulation models, and use the variables to estimate the model's parameters (operationalize the models).

Statistical tests will be conducted on the estimated models. There will be evaluations of user satisfaction with: (1) the speed, relevance, and quality of ad hoc query results; (2) the system interface; (3) model appropriateness; and (4) the quality of the system explanation. Simulations will be statistically tested for accuracy and confidence intervals will be established for the results. Tests will also be conducted on the system's ability to improve the decision-making maturity of the user.


Figure 6

In the testing, part of the data will be used to develop the system's pattern detection and explanatory models, and the rest of the data will be used to statistically test the efficacy of the system in predicting patterns of care, outcomes of care, and effects of cancer therapies. Figure 6 outlines the research design.

Back to Top

Specific hypotheses

During the project, the investigators will design and develop the information system to support cancer surveillance from health claims data. Then the investigators will test the efficacy of the system in predicting patterns of care, outcomes of care, and effects of cancer therapies. It is hypothesized that:

* The information system can detect important patterns in health claims data, such as practice patterns at various stages of treatment, end-of-life care patterns, long-term complication patterns for survivors, and diagnostic test use patterns.

* The information system can explain the underlying causes of the detected patterns, such as the factors that influence the care received at the end of life; the effect of carve-out contracts for cancer treatment on treatment patterns and outcomes; and environmental, biological, and hereditary causes of cancers.

* The information system can use predictive modeling to identify subpopulations that might have a higher incidence of cancer and who might consequently affect the cost structures of cancer care.

* The information system can improve the clinical and administrative practice of cancer surveillance.

In short, the evaluation plan can assess whether or not the proposed information system can help reduce the national cancer burden.

Back to Top

POTENTIAL BENEFITS

After the project is complete, interested parties will have a proven information system that will enable users to easily answer several research questions posed in PA-99-015, Cancer Surveillance Using Health Claims-Based Data System. 1 In particular, system users will be able to determine the following from the health claims data:

* whether claims data can be used to assess practice patterns for cancer directed treatment immediately following diagnosis and for long-term follow-up and treatment over the course of the disease

* the patterns of care for patients with specific cancers

* the pattern for end-of-life care and the factors that influence this care

* how treatment patterns and outcomes are affected by the use of carve-out contracts for cancer treatment

* the long-term complications for cancer survivors as sequelae to their cancer treatment

* whether claims data can be used to assess the use of diagnostic tests

* how diagnostic tests are used

* the emergence of new technologies for detecting, diagnosing, or treating cancer or precancerous conditions

* the relationship between the introduction of health care innovations and cancer

* whether claims data can contribute to the identification of environmental causes of cancer

* the degree to which claims data provide additional information than what is routinely collected by cancer registries

* whether claims data can be used to assess cancer recurrence or metastasis

* whether claims data can be used to augment the case ascertainment of cancer registries

* whether dental claims data provide useful supplementary data, especially as regards oral cancer detection or the prevention and treatment of oral complications of radiation or chemotherapy

Answers to these questions are especially timely in view of the interest expressed by NIH.

Several derivative benefits are expected as a result of this research effort. The project will identify data available only from health claims, barriers that must be overcome to add health claims data to registry data, and the costs and security issues involved in performing DSS-based cancer surveillance. Interested parties, armed with pertinent knowledge, will be in a better position to evaluate effectively the efficacy of specified cancer treatments and interventions.

The developed information system will provide a vehicle to efficiently perform the evaluations of cancer treatments and interventions before actual policies are put in place. Effectively designed treatments and interventions then should reduce the incidence, morbidity, mortality, and costs associated with the national cancer burden, especially as related to oral and pharyngeal cancers or the oral complications of cancer therapies. The developed information system also can serve as a standard for similar future health care analyses and evaluations. Based on the results from previous, related research, the potential cost savings can be expected to reach several billion dollars. 12

Back to Top

CONCLUSIONS

A variety of methodologies and technologies are integrated within, and delivered through, the proposed CSS concept. Specifically, the integrated support includes the following:

* **data warehousing** to develop methods to view health claims data, extract and capture elements pertinent to cancer surveillance, and warehouse the captured data for interested parties

* **data mining** to access the data warehouse and detect care, outcome, and therapy patterns from the captured health claims data

* **explanatory models** that rely on statistical models that explain the detected care, outcome, and therapy patterns in the health claims data

* **a decision support system** that readily delivers the developed data warehousing, data mining, exploratory, and predictive models to the interested parties

By utilizing this concept, these parties can readily: (1) organize relevant claims data, (2) detect cancer patterns in general and special populations, (3) formulate models that explain the patterns, and (4) evaluate the efficacy of specified treatments and interventions with the formulations.

While promising, the concept needs empirical testing and evaluation. The proposed evaluation plan offers one roadmap for this process. In particular, the plan offers: **utility analysis** to evaluate the utility of the analytical approaches and IT in assessing the national cancer burden for general and special populations and **a testing methodology** to demonstrate the utility with Medicaid data.

Back to Top

REFERENCES

1. National Institutes of Health, PA-99-015, *Cancer Surveillance Using Health Claims-Based Data System*. Washington, DC: National Cancer Institute, 1998. [Context Link]

2. Barquin, R., and Edelstein, H. *Planning and Designing the Data Warehouse*. Englewood Cliffs, NJ: Prentice-Hall, 1997. [Context Link]

3. Kimball, R. *The Data Warehouse Toolkit*. New York: John Wiley and Sons, 1996. [Context Link]

4. Agrawal, A., and Srikant, R. "Mining Sequential Patterns." In *Proceedings of the IEEE International Conference on Data Engineering*, 1995. [Context Link]

5. Fayaad, U., Piatetsky-Shapiro, G., and Smyth, P. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Communications of the ACM* 39, no. 11 (1996): 27-34. **Find It @MU** | [Context Link]

6. Berndt, D.J., and Clifford, J. "Finding Patterns in Time Series: A Dynamic Programming Approach." In *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, et al. Cambridge, MA: AAAI Press/MIT Press, 1996: 229-248. [Context Link]

7. Friedman, J.H. "Multivariate Adaptive Regression Splines." *Annals of Statistics* 19 (1989): 1-141. [Context Link]

8. Borok, L.S. "Data Mining: Sophisticated Forms of Managed Care Modeling Through Artificial Intelligence: Review." *Journal of Health Care Finance* 23, no. 3 (1997): 20-36. **Find It @MU** | **Bibliographic Links** | [Context Link]

9. Limb, P.R., and Meggs, G.J. "Data Mining-Tools and Techniques." *British Telecom Technology Journal* 12, no. 4 (1995): 32-41. **Find It @MU** | [Context Link]

10. Forgionne, G.A., and Kohli, R. "HMSS: A Management Support System for Concurrent Hospital Decision Making." *Decision Support Systems* 16 (1996): 209-229. **Find It @MU** | **Bibliographic Links** | [Context Link]

11. Forgionne, G.A., and Kohli, R. "A Model To Measure DSS Effectiveness: Theory and Empirical Analysis." *Journal of Decision Systems* 7, no. 2 (1998): 105-122. [Context Link]

12. Forgionne, G.A. "HADTS: A Decision Technology System To Support Army Housing Management." *European Journal of Operational Research* 97 (1997): 363-379. **Find It @MU** | [Context Link]

Key Words: breast cancer surveillance; data mining; data warehousing; decision support systems

---

IMAGE GALLERY

Select All

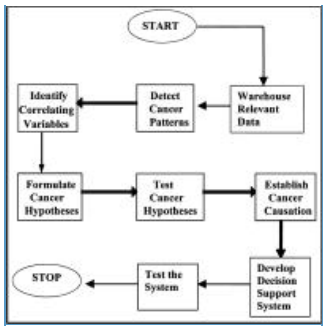

Figure 1

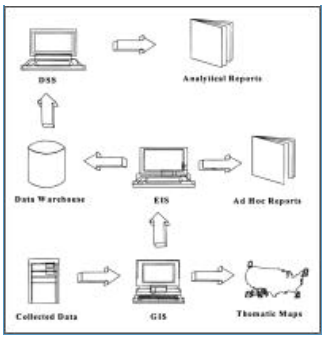

Figure 2

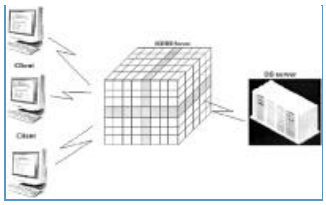

Figure 3

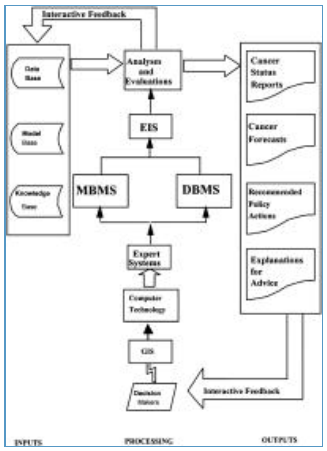

Figure 4

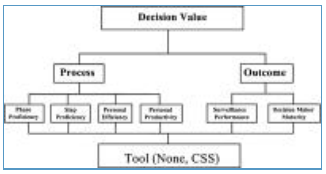

Figure 5

Table 1

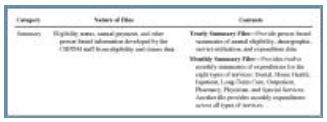

Table 1

Table 2

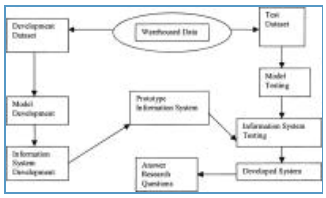

Figure 6

Back to Top

Previous Article | Table of Contents | Next Article