Health Centers and the Data Warehouse

Data warehouses may be relatively new to the healthcare provider environment, but they have been in use in other industries for many years, particularly in retail and manufacturing. Consider the amount of data that a large chain retailer can collect such as information about their customers, the products they purchase, how much they pay for those items and how they pay for them, in what stores the items are purchased, and when they are purchased. Link that to shipping and inventory data, human resource information and all other manner of data, and management has the tools to improve productivity and efficiency across the enterprise, negotiate better pricing with suppliers, thus lowering costs, and concurrently improve the quality of the customer experience.

With healthcare generating much more data than the retail industry, the opportunities to harness that data and then manage and manipulate it in such a way that it provides similar benefits to all organizational stakeholders are significant.

The use of a data warehouse:

♦ Supports strategic planning and quality management,
♦ Fosters improved outcomes for patients, populations, and the provider organizations
♦ Enables public health initiatives at the state and national levels,
♦ Permits healthcare providers to influence policy and payment reforms.
Because data warehouses require considerable resources to design, implement, and succeed, they are less likely to be found at individual community health centers, and more likely to be sponsored by either a health center-controlled network (HCCN), state primary care association (PCA), large multi-site health center, or other such group of provider entities. Achieving the intended benefits depends, however, on a number of factors not the least of which is the commitment of the participating organizations to openly share information and discuss and analyze its meaning together.

This issue brief describes:

♦ Characteristics of data warehousing
♦ Its usefulness in the context described
♦ Considerations for health center staff when developing and participating in a data warehouse.

WHAT IS A DATA WAREHOUSE?

A data warehouse is an extract of an organization’s data — often drawn from multiple sources — to facilitate analysis, reporting and strategic decision making. It contains only alpha-numeric data, not documents or other types of content. The data is stored separately from the organization’s primary applications and databases such as practice management systems and electronic health records. The data is transformed to match a uniform data model, cleansed of duplicates and inaccuracies, and is extracted with business intelligence and reporting tools. A data warehouse contains the entire scope of data and can be used for both very general and very specific analysis and reporting.

A data warehouse is distinct from a data mart, which is designed to focus on a single type of data or single business problem or report. Data marts are often subsets of data warehouses.

THE USEFULNESS OF A DATA WAREHOUSE

A data warehouse is an example of the journey that data takes, when combined with context, to become information. Prior to application of context it is just a collection of numbers and letters, bits and bytes. Yet information is still not enough to enable an organization to learn from and act based on what they have collected. Dutch physicist Heike Kamerlingh Onnes who won the Nobel Prize in 1913 for having discovered superconductivity, knew even then what that missing ingredient was to take information to the next step, as indicated by his motto inscribed on a sign posted at the entrance to his laboratory: “Through measurement comes knowledge.” The ability to use accurate data and timely information to objectively measure and, therefore, proactively manage clinical outcomes and business processes demonstrates the value of a data warehouse.

Certainly, there is an abundant amount of data to be found in:

♦ Practice management systems
♦ Electronic health records
♦ Financial applications
♦ Other systems and applications operating in the health center environment.

These systems are transactional in nature, however, and even the most advanced integrated practice management and electronic health record applications utilizing a single unified database do not have all of the information or tools necessary to support executive and strategic decision support. It is the opportunity to integrate and normalize data from multiple disparate sources, and to apply analytic tools and processes that leads to the usefulness of a data warehouse. Adding the ability to use the element of time to provide a longitudinal view across all of the data adds yet another dimension to its utility.

Data warehouses are generally most useful to a community health center within the context of participation in a larger organization such as a HCCN, large multi-site health center, or PCA. This is both because the fixed costs of establishing a warehouse when shared across a group of providers allows each to achieve economies of scale, and because the volume of data that can be collected and used for trending and comparison is that much greater. Aggregate provider warehouses are also important at the patient specific level because a particular patient may be seen in multiple participating centers.

### Usefulness for the Patient

A data warehouse provides the ability above and beyond an electronic health record to manage and improve the quality of care down to the individual patient level. This can be accomplished through a number of approaches, though all must start with the application of medical informatics. Medical informatics is the scientific discipline that determines what data is useful and necessary for healthcare practice, education, and research. Consider the potential to manage a patient with a chronic condition, knowing not only their personal history, but how other patients with similar characteristics respond to a particular therapy or treatment protocol. The benefit accrues not only from the use of this knowledge to guide the course of treatment, but from the availability of what is in essence a continuous feedback loop – a data repository continuously and repeatedly updated as new information becomes available.

### Usefulness for the Population

Managing the health of populations is fostered through the use of data warehouses. Screening initiatives and preventive care measures are important, if they work. But how do you know if they are, in fact, working? A data warehouse and the ability to analyze various pieces of data over the course of time will likely provide the answer. As with chronic disease management, the ability to drill down to the population and/or provider-specific levels can facilitate problem identification and resolution. The Health Disparities Collaboratives collected data for a similar purpose, but did not have access to all of the other information within a given organization, thus limiting usefulness as an analytical tool.

Public health needs can also be satisfied by a data warehouse, particularly in a model sponsored by a HCCN or PCA in which data from multiple sites is available. The ability, for example, to track and trend diagnoses and antibiotic prescribing support proactive bio-surveillance efforts as well as retrospective population focused analysis.

---

Combining information stored in a data warehouse with geographic information systems can provide a view of patient populations not only with the added dimension of time, but now with the dimension of geographic region as well. A health center could pose questions such as:

- Are the demographics of the health center or the community changing?
- Are your patients moving or being displaced?
- Are there environmental factors linked to a specific geographic location that are impacting your patient populations?

Such questions can be addressed by combining your data warehouse with geo-mapping tools to expand your knowledge about the population that you serve.

**Usefulness for the Provider**

Data warehouse information extends beyond the patient and population level to the provider dimension. Clinical leadership can use data warehouse functionality across populations to identify best practices in the treatment of a chronic condition by identifying those providers whose patients have better outcomes, as well as to identify those providers that may need reinforcement based on the less than optimal outcomes of their particular patients. By being able to drill down further to the patient level, greater understanding can be gained.

Why do Dr. Smith's patients have better outcomes than Dr. Jones' patients? Are they both achieving clinical goals? Perhaps the data warehouse can provide insight into not only the effectiveness of care, but also allow for examination of the patients' adherence to treatment plans and medication compliance. Maybe there are extenuating circumstances that apply to Dr. Jones' patients but not to those of Dr. Smith.

A data warehouse has the potential to:

- Become an investigative tool to discover opportunities to improve care and outcomes,
- Foster the “collegial competition” that many providers engage in when data and information are transparent,
- Improve outcomes,
- Help satisfy the goals of pay-for-performance initiatives.

**Usefulness for Organizations**

When considering the usefulness of a data warehouse at the organizational level, an organization can be defined as a health center, PCA, or an HCCN. All three kinds of organizations benefit from a data warehouse.

A data warehouse when combined with appropriate analytic tools provides business intelligence to:

- Use information to gain knowledge about an organization, how it is operating, and how to improve the decision making processes.
- Develop benchmarks to determine how individual centers within a given HCCN or PCA are performing and to establish standards against which they can be measured.
- Plan for operational needs such as space, supplies, and human resources.
- Provide insight for financial modeling.
- Demonstrate performance information when negotiating reimbursement levels.

For **individual health centers**, a data warehouse can:

1. Enhance reporting opportunities for internal management purposes as well as for grant and other regulatory requirements.
2. Provide information of use to quality information organizations.
3. Aid in fiscal management and projections.

---

National Association of Community Health Centers, Inc.®

September 2008
PARTICIPATING IN A DATA WAREHOUSE

Choosing to participate in a data warehouse initiative is not a matter to be taken lightly. Certainly the benefits are attractive, but there are issues to consider and address to ensure the investment yields tangible and useful results.

Privacy, Security and Secondary Usage

Issues of privacy are nothing new to healthcare, yet the aggregation of vast amounts of patient data found in a data warehouse – especially when combined with tools specifically designed to mine such data – presents even more challenges for participating organizations in their roles as stewards of the information.

Privacy

**Patient Identifiable Information**

– HIPAA regulations restrict the use and disclosure of patient identifiable information by health centers without patient authorization to outside entities (such as an HCCN). State privacy laws, which vary from jurisdiction to jurisdiction, may further restrict the use and disclosure of patient identifiable information. Moreover, legal responsibility for maintaining confidentiality and accounting for such uses and disclosures, and the potential liability if such laws are not complied with, falls upon the health centers participating in a data warehouse initiative. The fact that an HCCN may include health centers across state lines is an important consideration as privacy laws in the strictest participating state should then be observed in matters relating to the data warehouse to ensure compliance.

**Business Associate Agreement** – In most cases, information gleaned from a data warehouse is in aggregate form and unidentifiable. That the raw data associated with patient identifiers exists in the data warehouse does, however, result in the need to address privacy further. To disclose patient-level data to the HCCN, each covered entity must execute “business associate” agreements with the HCCN that specifies how the protected health information (PHI) may be used and disclosed. Furthermore, the HCCN may not disclose the PHI of one covered entity to another covered entity:

- Unless the PHI is disclosed for treatment purposes, or
- The covered entities both have or had a relationship with the patient whose PHI is at issue, or
- The PHI is disclosed for health care operations (including QA activities).

An exception to the disclosure requirements is if the covered entities meet the definition of an “organized health care arrangement” under HIPAA, which is not likely because independent health centers do not usually hold themselves out to the public as an integrated health care system.

For further information about how to manage privacy and security legal requirements, CHCs should review NACHC Information Bulletin “Strategies to Minimize Legal Risks Related to Health Information Exchange, Health Center Controlled Networks, No. 13 (Feb. 2008).

**Public Health Purposes** – In cases where data will be used for public health purposes, covered entities may only disclose PHI without patient authorization to legally authorized public health authorities and their agents for specified public health purposes. Even if the disclosure was for public health purposes, this would not be sufficient to authorize the covered entity to disclose PHI to the HCCN because the HCCN is unlikely to be a legally authorized public health authority.

**Research Purposes** – Should data be used for research purposes, it is important to remember that the term “research” has a distinct meaning under HIPAA that refers to a systematic investigation to develop or contribute generalizable knowledge. In general, PHI may not be disclosed for research purposes unless the patient has authorized the disclosure. However, when similar activities are undertaken by a health care provider for the purpose of improving treatment or the quality of care of the provider’s own patients, it is advisable, whenever possible, to consider whether such activities are “quality improvement activities” or “quality assurance activities” because such activities constitute health care operations...
and, as such, may involve the use and disclosure of PHI without patient authorization.

Security

While security often operationalizes privacy, particularly in defining and supporting access and authorization models for the data, it is also an important component in terms of maintaining the integrity of the data. A data warehouse is only useful if it is available, so security measures must address business continuity and disaster recovery planning as well.

- The HIPAA Privacy Rule requires covered entities to ensure appropriate administrative, technical and physical safeguards have been implemented to protect the privacy of PHI and to reasonably safeguard PHI from intentional or unintentional disclosure of PHI in violation of HIPAA.

- The HIPAA Security Rule requires covered entities to protect against reasonably anticipated threats to the security of electronic PHI ("ePHI"), protect against reasonably anticipated misuses or disclosures of PHI, ensure that their workforces comply with the Security rule, and to obtain assurances of confidentiality and security from their contractors, such as an HCCN, which may be accomplished through a business associate agreement.

Secondary Use of Information

Secondary use of information is also a growing concern. Patient identifiable information is ultimately owned by the patient and while informed consent may have been obtained initially, the usefulness of information contained in a data warehouse grows concurrently and proportionately with the length of time that that data is maintained and as the volume of data in the warehouse grows. Such vast amounts of information are not often found elsewhere and may be attractive to those who may wish to use it for other purposes. Not only might such secondary usage not have been approved by the patient initially, there are many ethical and moral issues regarding such use of data that have arisen in the recent past that may also have a bearing on how an individual may wish their data to be used. In addition,

- HIPAA does not generally permit disclosures of PHI without patient consent for research purposes.

- If the PHI is de-identified to produce a “limited data set”, the information may be disclosed without patient authorization.

Stewardship of the data on behalf of the patient is an important and serious consideration to be addressed by the participating organizations, which must also ensure that issues such as secondary usage are addressed in agreements with the HCCN, PCA or other vendor, so that the vendor is provided the legal parameters of how and when data may be shared or sold to outside entities.

Cultural Issues

It is important to realize that there are substantial cultural and organizational issues associated with participation in and development of a data warehouse. These issues generally fall into the categories of governance and transparency, or focus on technology rather than the cultural context of the organizations involved. None of these issues are independent of each other. The culture of a particular organization informs its attitudes toward governance and transparency issues, and these are reflected in the attitudes around data warehousing.

- Data warehousing requires a deep commitment to sharing information and to using that information in ways that may not appear to benefit any single unit or individual within the organization. An example would be the analysis of outcome and quality data across an HCCN to determine why certain provider’s outcomes are better than others (as described above). There are many real-world examples of healthcare organizations that have not wanted to share information across organizational lines.

- Issues of readiness and transparency must be addressed at the level of each individual healthcare organization committing...
to data warehousing and at a higher policy level at the HCCN, PCA, state and national levels.

These issues were nowhere as evident as in the Santa Barbara, California health information exchange project. Although the project did not involve CHCs, HCCNs or PCAs, and only peripherally involved data warehousing, it was one of the first large-scale multi-organizational projects involving the sharing of data to founder, due in part to its failure to adequately address such issues. Brailor’s3 detailed description of the lack of cultural focus in this project serves as a cautionary tale for anyone attempting data warehousing or health information exchange.

Cost and Resource Allocation Issues

There are significant costs and resource allocation considerations associated with the development of and participation in a data warehouse initiative; how those costs are divided among participating organizations will vary for each project.

Start-Up Costs – In many cases, start-up costs are supported by grants and other sources of funding.

Sustainability – Often neglected, costs and resource requirements associated with the use and maintenance of a data warehouse, growth in terms of the amount of storage necessary for increasing amounts of data and the addition of new participating organizations, and changing and more complex analytics needs must all be considered.

But how do organizations garner support to continue funding the initiative? Return on Investment (ROI) for healthcare information technology initiatives is often elusive at best due to a number of factors, not the least of which is that there are several variables that may impact any ROI model. Dale Sanders, a pioneer in enterprise data warehousing (EDW) for healthcare states, “In large part, the commitment to developing an EDW is a commitment to seeking a better understanding of your business and to continuous quality improvement. This is what makes predicting an EDW’s ROI so difficult – if the company culture is not committed to continuous quality improvement, the best EDW technological solution in the world will still have a negative ROI.”

DEVELOPING A DATA WAREHOUSE

The development of a data warehouse is a huge undertaking, and requires a considerable commitment of time and effort as well as dollars. Experience has shown that development:

♦ Is often a multi-year project.
♦ Requires substantial planning, infrastructure development, programming and testing before a data warehouse can be consistently used.
♦ Involves many decisions about technical and non-technical issues, responsibility for data security, integrity and ownership, governance and usage.
♦ An organization must be prepared not only to make these decisions initially, but to continue to be responsible for the implementation, use, and maintenance of the data warehouse into the future. But just as the data itself may grow over time, so will the value. Consequently, the initial benefit may be limited but will grow in a manner commensurate with the volume of data and experience with accessing it. Health centers need to consider that the benefits of data warehousing are best achieved over a longer rather than shorter horizon.

---


The major elements of developing a data warehouse include:

- Planning
- Designing and deployment of hardware and software infrastructure
- Testing

Planning the Data Warehouse

Planning is the most important aspect of implementation as the decisions made during the planning process will determine the success or failure of the project. Important components of planning include:

- Define the goals of implementing and using the data warehouse.
- Delineate the specific roles and responsibilities of both participating organizations and individuals.
- Develop a governance model for the data warehouse which will clearly identify who will be responsible for each aspect of its operation and use.
- Determine a data ownership model.
- Develop a business model.
- Describe the model for use and sharing of data and analytic results (internally, among, and external to the participating organizations).

Perhaps the most important aspect of planning is:

- Ensure adequate and appropriate representation and participation from each of the participating organizations, specifically including users and key stakeholders as well as information technology specialists.

Users and key stakeholders will facilitate making sure that cultural issues, if not addressed, will at least be visible during the planning process.

Designing the Data Model

Data Integrity

Data integrity is perhaps the most critical success factor for a successful data warehouse. Data available for analysis and reporting must be transformed so that it conforms to the established data model, and it must be cleansed so that it is free of duplicates and ambiguous or incorrect information. This is accomplished through the use of an Extract, Transform and Load (ETL) tool (or set of tools). ETL is an application that operates against a set of definitions to normalize the data. For example, identical data reported in different formats are mapped to the same form (Doctor, Dr., Dr., dr. and dr are all mapped to Dr). Such mapping is usually performed using temporary or staging tables.

Transformed data in staging tables is mapped into a data model schema, and stored in the data warehouse tables.

The data warehouse tables are optimized for retrieval, analysis and reporting so the schema used is often quite different from the database table design of the system from which the data originated.

Design of the Schema

There are two approaches to designing this schema – top-down and bottom-up. Each have benefits and drawbacks.

Top-Down models for data warehouses – are designed so that individual facts (e.g. cost, procedure, diagnosis, facility) are structured by different dimensions. Cost, for instance, could be structured along a time dimension; cost per day, cost per week and so on. In a top-down design, all the different dimensions and associations of cost (and every other type of data to be included in analysis) would be defined and represented in the schema. This is a large and complicated effort that requires a schema design specialist and very detailed specifications of how the data is to be used.
**Bottom-up design** – creates one or a series of data marts to address specific business or clinical problems and then attempts to unify them into a single data model and data schema. This is usually quite difficult, and often the data marts designed for different purposes are loaded and stored separately. The primary disadvantage in this case is that large-scale analysis or reporting across data marts are difficult if not impossible to achieve. Bottom-up design can work, and it is often the most cost-efficient way to start a data warehousing effort, but unification of individual data marts requires a top-down schema be designed.

**Data Dictionary** – regardless of which method is used, a data dictionary that defines the organization of the data in the warehouse and documents its contents is also essential. All commercial database products have data dictionary components. It is convenient as part of the schema design process to address additional elements that you might want in the data dictionary for your analytic data. These might include relationships among specific data elements (to enable or optimize certain types of analysis or reporting), information on the origin of the data, information on the transformation process and others. The data dictionary makes it possible for analytic and reporting applications to more correctly use the data that is available.

**Validity of Data**

Once the schema is designed, and the transformation rules are written, the validity of the data must be addressed. A set of rules are written to check the transformed data for duplicates and for any obvious errors, contradictions or inaccuracies as far as can be anticipated. These rules are applied to the staging tables (that is transformed data) as part of the data model loading process. They can not be applied before the data is transformed, as it will not be possible to tell if data is duplicated or inaccurate until it is in a standardized form. The development of these rules also requires effort and usually a specialist programmer.

**Technology for the Data Warehouse**

It is essential to understand that a data warehousing project is not a technology project, even though it seems that the technology aspects of such projects may be overwhelming. Unless a data warehouse fits into the business and clinical goals of a healthcare organization, it can not be a success. Once it is decided to develop a data warehouse and the organization has completed the planning and data design, the organization must ensure that its technology infrastructure is adequate to support this venture.

- The first technological necessity is a reliable **high-speed network** (100 MBPS minimum, 1 GBPS preferred) throughout and between the participating organizations. All internal network components including switches and routers must be able to support this network. The network is used not only to transfer data and results, but to maintain the infrastructure itself.

- A set of **servers** will be required to support each aspect of the architecture of the data warehouse. These should be as modern as possible, (e.g. running at least Microsoft Server 2003 R2 or Microsoft Server 2008, with 2-4 GB of virtual memory) and have associated reliable and scalable disk storage to match the planned size of the data warehouse plus 30% allocated for growth. A data warehouse will usually require multiple stand-alone servers. One will be used to support the selected relational database (Oracle, SQL Server, DB2 etc.) that stores the data from the health center’s practice management, EHR and other operational applications, and one to support the ETL application and data warehouse store (also often a commercial relational database). Often, an additional server is needed to support sophisticated business intelligence, statistical analysis or modeling applications.
An infrastructure such as the one described requires information technology staff resources to support and maintain it.

A portfolio of software applications will also be required including, but not limited to: a commercial relational database (or other storage system), analysis and reporting tools, business intelligence or standalone analysis applications (e.g. Cognos, Business Objects, SAS), and an ETL application. Figure 1 below depicts the components of the warehouse and the relationships between them.

Developing a data warehouse is real work. It requires significant planning which can sometimes take as much as 9-12 months. It requires a good deal of programming to develop and write rules for data extraction and transformation as well as data integrity. It requires the design of a data model and its implementation as a data schema. It also requires the development of a network, hardware and software infrastructure and the continual redesign and maintenance of that infrastructure.

Considering Non-Technical Essentials

Perhaps even more difficult are the non-technical issues that are inevitably part of the design and implementation process including the development of and commitment to goals, a long-term business model and funding process. This requires commitment at every level of each of the involved organizations as well as executive attention if it is to succeed. A data warehouse effort at the HCCN or PCA level, however, can provide deep insight into both the business and clinical operations that are the lifeblood of these organizations.

Figure 1. Generic Data Warehouse, Functional Architecture (in part from Verma, 2001. op.cit.)
WHAT THE FUTURE HOLDS

In the words of Yogi Berra, “Prediction is hard, especially when it’s about the future.” That being said, we are comfortable with the following speculations. Two technologies are emerging as important parts of near-future data warehouses – search and visualization.

Search – The convenient and powerful means for locating specific information in a set of data has become one of the primary ways that people not only accomplish their work but also organize their work. Search will become more and more deeply integrated into large-scale data storage such as data warehouses so that it can be applied to the transformed data that make up the warehouse; to the analytic results that are stored; and to reports that are generated. This will enable a deeper and more effective use of the data and analysis.

New Forms of Visualization – The manner in which data is presented for interpretation will make analysis much more productive. We better understand analytic results, and especially the relationships among results, if we actually see a visual representation of those relationships rather than the statistics that represent the relationship. Every business intelligence and analytic package associated with data warehouses will offer a wide variety of visualization capability that, over time, will be easier to use and interpret. Visualization is already changing how analysis is done and interpreted in many other fields, and healthcare is no exception.

As health information exchange initiatives continue to grow at the local, regional, state and national levels, there will be increasing opportunities to leverage these initiatives to aggregate clinical and demographic data from across multiple entities. The ability to integrate data from such initiatives with a data warehouse will provide large-scale research and strategy opportunities and this is where much of the benefit lies with a data warehouse.

New Kinds of Organizations – The use of data warehousing by a CHC, HCCN or PCA provides a wide variety of opportunities to understand and improve both the business and clinical operations of the organization. As more data becomes available, and more analysis is done, the opportunity for new types of organizations, some actual and some virtual, becomes possible. These organizations could be based strictly on business characteristics of the associated organizations or on clinical characteristics of the patients served. Data warehouses and their related applications also allow for predictive modeling at a number of levels, be it space and resource planning, identifying new service lines, and financial planning. Such opportunities come as a direct result of an organization, network, or PCA deepening its understanding of its operations through the development and evolution of analysis and reporting based on a data warehouse.

This technology and the organizational and operational improvements that come as a result of its use have the potential to change CHCs, HCCNs and PCAs substantially for the better.

Data warehouses and their related applications also allow for predictive modeling at a number of levels, be it space and resource planning, identifying new service lines, and financial planning.