Chapter 2
Clinical Research Systems and Integration with Medical Systems

Joyce C. Niland and Layla Rouse

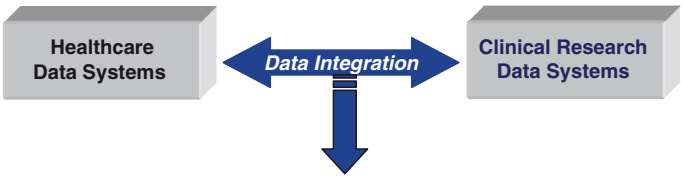
Abstract Integration of the Electronic Medical Records (EMR) with clinical research systems has the potential to greatly enhance the efficiency, speed, and safety of cancer research. New hypotheses could be generated through mining of EMR data, observational studies may be conducted more rapidly, and clinical trial recruitment and conduct could be greatly facilitated. Such enhancements will be accomplished through secondary use of EMR data for research and the development of automated decision support systems that rely on EMR data. In this chapter, we define the various types of EMR and clinical research data systems in use and describe the goals and rationale for integrating these two types of systems to enhance research as well as quality of care. The various approaches and benefits to integrating EMR and clinical research systems are discussed. While major benefits are conferred by such system integration, many challenges exist as well, such as the need for stringent data quality assurance, appropriate granularity, metadata and person index management, and extremely careful handling of data access and security issues. Furthermore, the movement toward the EMR within the USA has been slow to date, hampering these data integration efforts. However, recent legislation to incentivize the adoption of EMRs will make the feasibility and utility of EMR data integration to support clinical research more promising in the near future.

2.1 Introduction

It is critical that the efficiency, speed, and safety of cancer research be continually enhanced to make more rapid inroads and progress in battling this devastating disease. One approach to achieving these goals is to ensure that when conducting clinical research, full advantage is taken of the emerging role of electronic medical records (EMRs) in the field of cancer care. Yet an aspect of EMRs that has received

J.C. Niland (✉)
City of Hope National Medical Center, 1500 East Duarte Road, Duarte, CA 91010, USA
e-mail: JNiland@coh.org

M.F. Ochs et al. (eds.), *Biomedical Informatics for Cancer Research*,
DOI 10.1007/978-1-4419-5714-6_2, © Springer Science+Business Media, LLC 2010



	Secondary Use of Data	*Automated Decision Support*
I. ***Speed***	- Identification of subjects for trial - Import of pre-existing data into e-forms	- Algorithms to cull out ineligible subjects - Metadata-driven mouseover information
II. ***Efficiency***	- Hypothesis generation via data mining - Prevalence data for study planning	- Pattern dectection rules applied to data - Graphical contour plots of prevalence
III. ***Safety***	- Import of data to identify toxicities - Use of data to develop care guidelines	- Automatic grading of adverse events - Rule-based guideline adherence data

Fig. 2.1 Synergies between clinical research and medical data systems

little attention to date is the potential benefit of these systems to clinical research (Powell and Buchan 2005).

The integration of EMRs with clinical research systems enables two key forms of functionality: *secondary use of data* and *automated decision support*. Through the former, integration of these two types of data systems can facilitate the efficiency and speed with which cancer clinical research can be conducted. Through the latter, such integration can greatly improve patient safety, as well as efficiency, as clinical research is being conducted. The synergistic nature of these systems and the goals of each are depicted in Fig. 2.1. In this chapter, we will discuss the approaches, benefits, and challenges of integrating clinical research systems with medical care systems. First we introduce and define the terms and processes that will frame our discussion.

2.2 Electronic Systems to be Integrated

2.2.1 Clinical Research Data Systems

Clinical research data systems take on several different forms and functions. One of the most frequently deployed clinical research systems can be defined as a Clinical Data Management System (CDMS) which is used in clinical research to

manage the data of a clinical trial (i.e., an experimental interventional study conducted with human subjects), as well as other forms of clinical research such as observational, outcomes, or epidemiological trials (Summerhayes 2002; Tai and Seldrup 2000; Greenes et al. 1969; Clinical Data Management System Wikipedia 2009). The data to be stored in the CDMS may be gathered on paper forms, such as Case Report Forms (CRFs) in the case of a clinical trial, or on survey forms, questionnaires, and other data capture forms for observational research studies.

Another form of clinical research system more specific to the area of interventional clinical trials is known as a Clinical Trial Management System (CTMS). A CTMS consists of a customizable software system to manage large amounts of data involved with the operation of a clinical trial (Choi et al. 2005; Payne et al. 2003; see Chaps. 10–12). Such a system not only provides a data capture interface and data storage, but also provides additional functionality, such as maintaining and managing the clinical trial planning, preparation, performance; tracking deadlines, data expectations, and milestones; and reporting of clinical trials for regulatory and analysis purposes. Modules for handling trial budgeting and patient study calendars may be included in the CTMS as well. Compatibility with other data management systems is a highly desirable feature of any CTMS or related study management software tool.

Clinical research data collected during the investigation of a new drug or medical device is collected by physicians, nurses, and research study coordinators in medical settings (offices, hospitals, and universities) throughout the world. Historically, this information was collected on paper forms, which were then sent to the research sponsor (e.g., a pharmaceutical company) for entry into a database and subsequent statistical analysis. However, this process has a number of shortcomings, including that data are copied multiple times, producing errors that may not be caught until weeks later. To alleviate such issues, another type of clinical research system that has evolved within biomedical research is known as a Remote Data Entry (RDE) system (Electronic Data Capture Wikipedia 2009).

RDE systems allow research staff to enter data directly at the medical setting, particularly useful when a multicentered study is being conducted with many institutions participating. By moving data entry directly into the clinic or other facility, data checks can be implemented during data entry, preventing some errors altogether and immediately prompting for resolution of suspicious entries. Early RDE systems often used "thick-client" software installed on a laptop computer, such that the system needed to be deployed, installed, and supported locally at every participating site. This process becomes quite expensive for the study sponsor and complicated for the research staff. For Cancer Centers that typically participate in many research studies simultaneously, this deployment model for RDE results in a proliferation of different systems being installed, leading to complexity for the users along with space constraints.

In recent times, the user interface for RDE has shifted to Web-based deployments, for entry of data by the research team member directly into the system. EDC systems do not require local installation initially or with each software upgrade, but rather can be deployed centrally by the study sponsor for immediate and seamless

access by users. Although these systems are better than thick-client approaches, there are still cross-browser dependencies that need to be dealt with to make these Web-based systems truly universal. Typically an EDC system will include not only the graphical user interface (GUI) component for data entry, but also imbedded validation algorithms to rapidly check data for errors or suspicious entries and a reporting tool for synthesis and display of the collected data (Electronic Data Capture Wikipedia 2009). Such functionality formerly would be made available as separate software solutions within the CDMS or CTMS; however, integrated end-to-end solutions are evolving more recently. While EDC systems are primarily designed for the collection of data for clinical trials, there is no prohibition for this type of system to become equally popular and useful for observational research studies as well.

The term "electronic data capture" also may encompass several types of technology, beyond an electronic replacement for the CRFs that are completed at the enrolling site (Handleman 2005). EDC systems can include data capture technologies such as interactive voice response (IVR) systems, for example, to allow patients to report information over the phone (e.g., "press a key from 1 to 5 to describe your current pain level, with 5 being the highest"). Patient-reported outcomes collected via electronic diaries, for example using a Personal Digital Assistant (PDA) such as a Palm Pilot or similar device to record information best captured at home, also may be considered a form of EDC (Handleman 2005).

For simplicity, within this chapter we will use the more global term of "CDMS" to encompass any form of electronic clinical research data system to be integrated with medical systems.

2.2.2 Electronic Healthcare Data Systems

There are many limitations of paper medical records, including unavailability at the point-of-care (a given medical record cannot be in multiple places at once), inconsistent legibility, duplication of information, poor indexing of information, and inconsistency of information (Winkelman and Leonard 2004). To help alleviate such deficiencies, electronic healthcare data systems have been evolving. The National Cancer Institute (2009) defines an Electronic Medical Record (EMR) as "a collection of a patient's medical information in a digital (electronic) form that can be viewed on a computer and easily shared by people taking care of the patient." Though often used interchangeably, the terms EMR and Electronic Health Record (EHR) have different meanings in medical informatics. An EHR is defined as a "a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting; including information on patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports" by the Health Information and Management System (Electronic Health Record Wikipedia 2009). While increasing familiarity with the term "EHR" is being engendered by the 2009

Health Information Technology for Economic and Clinical Health (HITECH) Act (see below), we will use more technical informatics term of EMR for purposes of this chapter.

A related but distinct form of electronic system for the capture, management, and reporting of health information is the Personalized Health Record (PHR), defined as an electronic system to allow individuals to enter and manage their own private health information. Because the data come directly from the person him/herself, the advantages are that the information may be more completely and accurately captured from a personal view point. However, a disadvantage is that lay persons may not fully comprehend or enter data that is fully correct medically. Generally the term Health Information System (HIS) is reserved for electronic systems that go beyond even the EMR/EHR functionality to include features such as automated decision support (see below), alerting, and/or lifetime cumulative records. Another term that may be encountered is The Medical Record (TMR), designed to be a truly comprehensive personal health record, including a birth-to-death, time-oriented database of all parameters related to a person's well-being. Integrating data from all points of delivery and from all medical specialties, the TMR is envisioned to create a historical view of the health-related course of events in a person's life (Hammond et al. 1997).

Again for simplicity within this chapter, we will use the term "EMR" to encompass the several types of electronic healthcare systems defined above that could potentially be integrated with clinical research data systems.

To be considered a "full" EMR, typically a minimum of three functional components must be included in the system: computerized physician order entry (CPOE), both for computerized prescription orders and orders for tests; reporting of test results; and capture of caregiver notes (Electronic Health Record Wikipedia 2009). One of the largest national EMR projects has been implemented by the United Kingdom's National Health Service (NHS) that will include 60,000,000 patients within a centralized EMR by 2010 (Electronic Health Record Wikipedia 2009). As another example, Alberta Province in Canada has deployed Alberta Netcare, a large-scale operational EMR system (Electronic Health Record Wikipedia 2009). The United States (US) Department of Veterans Affairs has deployed the largest enterprise-wide health information system that includes an EMR, the Veterans Health Information Systems and Technology Architecture (VistA) (Electronic Health Record Wikipedia 2009). This system allows healthcare providers to review and update a patient's EMR at any of the more than 1,000 VA facilities around the country. The New York City Health and Hospitals Corporation, serving over 1.3 million patients in the largest urban US healthcare agency, is another positive example of a successfully implemented EMR (Electronic Health Record Wikipedia 2009).

The National Center for Health Statistics (2006) has indicated that the overall adoption of EMRs has been slow within the USA, in spite of a study showing revenue gains after implementation of a new billing technology. US healthcare industry spends only 2% of gross revenues on information technology compared to upwards of 10% within other information intensive industries such as finance (CDC

National Center for Health Statistics 2006). If all medical payment transactions were handled electronically, it has been estimated that America could save $11 billion annually (Medicare Part B Imaging Services 2008). Yet, the vast majority of healthcare transactions in the USA still take place on paper. Data from the 2005 National Ambulatory Medical Care Survey indicated that only about 25% of office-based physicians reported using EMRs. While this represented a 31% increase from the 18% reporting use of such systems in 2001, only 9.3% of the responding physicians reported having a "complete" EMR in place as of 2005 (CDC National Center for Health Statistics 2006).

Beginning in 2005, a private nonprofit branch of the US Department of Health and Human Services, the Certification Commission for Healthcare Information Technology (CCHIT), was established and charged with developing a set of EMR standards, in order to certify vendors who are able to meet these standards. Hopefully such product certification will provide US physicians and hospitals with the mandate and justification needed to make the significant investment of EMR implementation. By July 2006, CCHIT had released its first list of 22 certified ambulatory EMR products, and starting in early 2007, EMR vendors began utilizing these certification criteria in building their systems (Certification Commission for Health Information Technology 2009; Certification Commission for Healthcare Information Technology Wikipedia 2009). Additional barriers to adopting an EMR, beyond the daunting cost, include the complexity of such systems and the necessary change management and training to allow widespread adoption. Furthermore, the lack of a national standard for interoperability among competing software options is a major hindrance to widespread adoption of such tools (National Archives and Records Administration 2008).

In 2009, President Obama signed into law an economic stimulus package known as the "HITECH Act": Medicare and Medicaid Health Information Technology; Title IV of the American Recovery and Reinvestment Act. One aim of this legislation is to incentivize more medical practices to implement EMRs, by providing a financial subsidy for physicians who adopt and meaningfully use certified systems. Using a "carrot and stick" approach, the bill also progressively reduces Medicare reimbursement to any physicians who have not implemented an EMR by 2015 (Health Information Technology for Economic and Clinical Health Act 2009; Center for Medicare and Medicaid Services Fact Sheet 2009).

2.3 Goals to be Achieved Through CDMS-EMR System Integration

2.3.1 Secondary Use of Data

A major goal in integrating clinical research systems with electronic healthcare data systems is to achieve "secondary data use." Safran et al. (2007) documented that secondary use of data can be defined as "non-direct care use of

personal health information (PHI), including but not limited to use of such data for analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities" (Safran et al. 2007). The first few uses listed above touch on this important intersection of clinical care and biomedical research. Individuals and organizations involved in cancer research that may benefit from secondary data use from medical records include health services researchers and clinical investigators, disease registries, health data organizations, healthcare technology developers, and research or policy centers (Anonymous 1993).

The Institute of Medicine also has identified that two types of patient records exist, emphasizing that all users should not have access to all parts of patient records, so that patient confidentiality can be maintained (Institute of Medicine 1991):

(a) *Primary records* are those used by healthcare professionals while providing patient care services to review previously recorded data or to document their own observations, actions, or instructions.
(b) *Secondary records* are derived from primary records and contain data elements to aid nonclinical users in supporting, evaluating, or advancing patient care.

Such secondary record usage includes biomedical research to advance the evaluation and discovery of new treatments, better methods of diagnosis and detection, and prevention of symptoms and recurrences. Cancer clinical trial research can be enhanced and informed by some of the data collected during the practice of care, such as comorbid conditions, staging and diagnosis, treatments received, recurrence of cancer, and vital status and cause of death. Analytic observational studies may involve the use of valuable standard of care data available in the EMR from the routine practice of medicine. Quality/safety measures can be gleaned from the EMR in support of outcomes and comparative effectiveness research to determine whether new clinical trial findings are being adopted into the community of all cancer patients, what the most effective strategies are in the general cancer population, identify population groups, and conduct epidemiological studies.

Secondary data on health-related subjects extends beyond only clinical medical information and may also include administrative records; statistical reports of governments and other agencies; political/legal documents such as voting records, wills, contracts, laws, and statutes; organizational minutes; proceedings and reports; poll returns; survey data; commercial, industrial, and institutional records; historical documents; personal documents such as letters; and communications in the mass media (Brown and Semradek 1992). While several of these data types can be instrumental in supporting various forms of research (e.g., epidemiological investigations into disease etiology, case–control studies with neighborhood controls matched on socioeconomic factors), for the purposes of this chapter we will restrict our discussion of research uses of data to the clinical information arising within EMR systems.

2.3.2 *Automated Decision Support Systems*

Another potential benefit to clinical cancer research that can be conferred by integrating the CDMS with the EMR is automated decision support to enhance patient safety and study conduct efficiency. Automated Decision Support System (ADSS) can be defined as a rule-based system that is able to automatically provide solutions to repetitive management problems (Turban et al. 1997). Software components of an ADSS include rules engines, mathematical and statistical algorithms, and workflow applications. While the ADSS is frequently found in business settings, such systems can play a crucial role in the continual struggle to improve the quality and efficiency of patient care. A healthcare ADSS is based on rules or algorithms that trigger an automatic decision; however, unlike in business informatics, such rules typically are not automatically acted upon without final review and acceptance by the medical caregiver to provide the human interaction, adjudication, and expert knowledge layer needed for safety reasons.

An ADSS is most useful in situations that require solutions to repetitive problems that mostly involve electronically available information (Automated Decision Support System Wikipedia 2009). For the ADSS to be useful, the problem situation at hand must be clear and well understood, and the required knowledge and relevant decision criteria must be very clearly defined and structured, requirements that are particularly challenging to achieve in the medical field. Particularly in the conduct of interventional clinical trials, and to some extent within observational research, the healthcare ADSS can be important for improving the safety and efficiency of clinical research.

2.4 Rationale for Integrating the CDMS with the EMR

An ideal solution for leveraging the EMR to support clinical cancer research would be to extract patient data directly from the EMRs, as opposed to collecting the data in a separate data collection software application (Electronic Medical Record Wikipedia 2009). The convergence between patient care EMR systems within the broader healthcare ecosystem is expected to continue and perhaps could one day reach the point where separate CDMS and EMR systems would not be needed. However, in today's world this combined usage of a single electronic system to fully serve both patient care and clinical research needs is not yet tenable and would be extremely challenging on several levels.

First, both EMRs and CDMSs represent "transactional" database systems, built to support a specific business process and set of use cases. Medical records are structured primarily for the clinicians and administrators (Electronic Medical Record Wikipedia 2009). An EMR is a dynamic entity, affording greater efficiency and quality control to the work processes of clinicians by providing data entry at the point of care, logistical information access capabilities, efficient information

retrieval, user friendliness, reliability, information security, and a capability for expansion as needs arise (Electronic Medical Record Wikipedia 2009).

Patient care systems can streamline the many daily interactions with thousands of patients to avoid slowing the healthcare process while using an EMR. Because they resemble paper-based formats, these highly structured data formats encourage a greater standardization of data entry, thus, promoting collaborative and goal-directed treatment planning (Stam and van Ginneken 1995). Within EMR systems, structured entries (e.g., codes, classifications, and nomenclatures) are more frequently used over paper-based records (Thiru et al. 2003). However, much of the patient care information still is not entered using close-ended standardized coding schemas, as is needed for research and data analytic purposes.

In addition, the transactional data records of an EMR are indexed by patient and often by account/visit numbers, unlike the need to index by protocol and subject within research data systems. Further, the large research data queries that need to be conducted could greatly impede the daily performance of the EMR and interfere with patient care, if performed directly within these healthcare-driven systems.

Therefore, given the current state of EMRs, the varied complexity of patient care vs. clinical research, and the different nature of the transactional databases that support the two processes, this convergence into a single shared-purpose electronic data system is not yet on the horizon. Instead at this juncture, the goals of secondary use of EMR data and automated decision support for clinical cancer research can best be achieved through the integration of EMR and CDMS data systems. The potential approaches to patient care–clinical research data integration, along with the many benefits conferred and challenges faced, are discussed in the following sections.

2.5 Approaches to Integrating CDMS and EMR Systems

2.5.1 Point-to-Point Data System Integration

One technical approach to integrating an EMR with the CDMS in order to support clinical research would be a "point-to-point" data integration solution. In this instance, the exported data from the EMR would be directly imported into the CDMS, most often as a scheduled "batch" file update, for example, nightly. First, a detailed systems analysis needs to be conducted to determine what data elements exist within the EMR that would be useful for research purposes and that exist in an appropriate form. Ideally, the data would be in coded or numeric format (e.g., M = Male, F = Female, numeric laboratory data results, etc.) and at an appropriate level of granularity or specificity to suit the research purpose at hand. While open-ended text-based data could be imported into the CDMS as well, this form of unstructured data requires substantial manual curation on the clinical research side before it could be readily used for research purposes. An intermediate level of data between coded and open text would be structured text, for example, arising from

physicians conducting dictations using formatted standardized templates, so that consistent information is obtained with each dictation, in a predictable order.

When only two systems are involved, a single EMR and single CDMS, the point-to-point data technical integration approach would be quite reasonable. However, more frequently there are several source systems that could provide data to support research, for example, financial systems for cost–benefit analyses, ancillary healthcare systems not linked into the EMR, etc. When more than the two systems are involved, point-to-point data integration solutions quickly begin to break down, and it becomes intractable to manage the numerous interfaces and synchronization of data across all systems. Integrated biomedical data not only enhances clinical research, but could also benefit hospital quality assurance, accreditation reporting, caseload and volume analyses, as well as genotype–phenotype correlative research if the "omics" forms of data are integrated as well. Therefore, a much more flexible scalable technical approach to this data integration problem is the data warehouse, as described in Sect. 2.5.2.

2.5.2 Data Warehousing

As shown in Fig. 2.2, the data warehousing approach to data integration, while challenging, provides a highly extensible, large dimension data integration solution (see Chap. 3). In this approach, there can be many "feeder" data systems that provide valuable source data to be exported to and stored in the data warehouse. These systems could include ancillary care systems (laboratory, pathology, etc.) that may pass through the EMR itself to the warehouse or may represent stand-alone data systems that pass data into the warehouse.

Additional source systems could consist of the observational and/or clinical trial data systems into which data are collected specific to research, that are not available through the patient care systems. Such data might include graded adverse events, best response to treatment according to the protocol definition, and outside medical care records pertinent to the research project, but existing only on paper and not coded in the internal EMR system. In addition, the "omics" data arising from genomics and/or proteomics experiments, and stored in systems such as those described in Chaps. 13 and 14, could be synthesized and imported into the warehouse in an aggregated reduced-dimensionality format, to be merged with the treatment and biological "phenomic" data on the patients. As with the point-to-point solution above, a detailed systems analysis and data dictionary (i.e., metadata, data defining data) development is a critical prerequisite to a successful data integration project such as data warehousing.

The process of extracting the specific subset of data from the source systems into the data warehouse is called the "Extract-Transform-Load" or ETL process (Adelman and Moss 2000). Via an automated, scheduled routine, the required data elements are exported from the feeder systems, typically nightly or weekly, transformed to meet the data model of the warehouse, and loaded into the data warehouse data

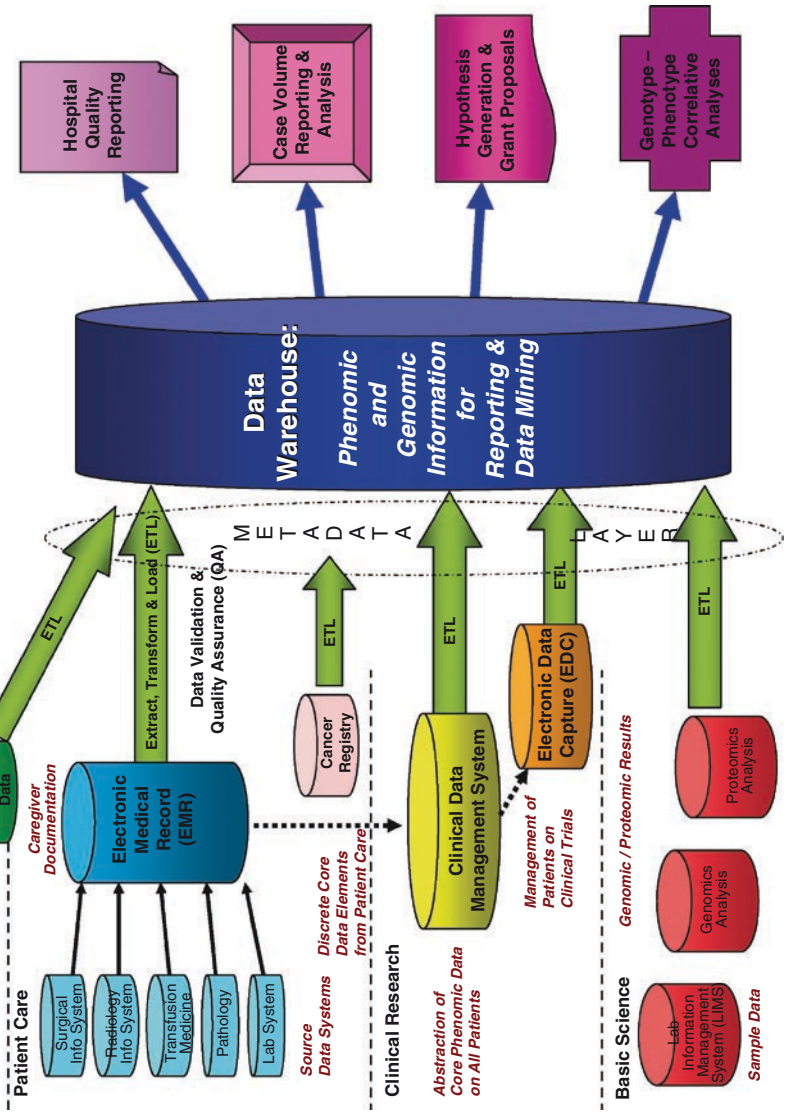


Fig. 2.2 Data warehouse schematic overview

structure. The underlying data model is usually specified as a "star schema" in order to provide the most efficient storage mode for data integration across the sources and subsequent abstraction for data mining purposes (Gray and Watson 1998).

As also shown in the Fig. 2.2, regardless of the technical integration solution, data quality assurance and validation are critical, as is metadata management, as described below. Once data are populated and integrated through a data warehousing approach, several types of "data marts" or subsets can be spun off from the main data store to meet different analytic and reporting purposes. These might include hospital quality assurance reports, evaluating whether complications of care and comorbidities are within acceptable ranges or case volume analyses to determine trends and plan for hospital beds and staffing. On the research side, clinical trials and observational research can be greatly facilitated through the integrated data, and genomic–phenomic correlative research facilitated through this highly valuable integrated data store.

2.5.3 Utilization of Standards

Regardless of which technical approach to data integration is utilized, it is crucial to follow existing and emerging data standards to ensure high-quality results and the ability to integrate across institutions, organizations, pathways, and diseases. Only through such standards will clinical research be advanced in a rapid highly organized manner, along with multicenter studies that are required to make more rapid biomedical discoveries.

Although few standards exist today for EMR systems as a whole, a number of standards exist relating to specific aspects of the EMR (Electronic Medical Record Wikipedia 2009). Adoption of several of these standards would greatly enhance the ability to conduct research on a global multidisciplinary scale when integrating data from the EMR for research. For example, the American Society for Testing and Materials (ASTM) International Continuity of Care Record (CCR) is a patient health summary standard based upon XML. The CCR can be created, read, and interpreted by various EMR systems, allowing easy interoperability between otherwise disparate entities (Electronic Medical Record Wikipedia 2009).

Standards for billing and financial purposes are available to potentially enhance data compatibility for research purposes, particularly because of their mandatory nature. The ANSI ASC X12 (EDI), a set of transaction protocols used for transmitting virtually any aspect of patient data, is in use in the USA for transmitting billing information, particularly as several of the transactions are required by the Health Insurance Portability and Accountability Act (HIPAA) (American National Standards Institute Accredited Standards Committee X12 Wikipedia 2009; Accredited Standards Committee X12 2009; Health Information Privacy 2009; Health Insurance Portability and Accountability Act 2009). Digital Imaging and Communications in Medicine (DICOM) standards are in widespread use for representing and communicating radiology images and reporting (Digital Imaging and Communications in Medicine Wikipedia 2009).

Interoperability can be defined as the ability of different information technology systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged (Electronic Medical Record Wikipedia 2009). The Health Level 7 (HL7) messaging standard is in use for interoperability among data from hospital, physician, EMR, and practice management systems (Health Level Seven 2009; Health Level 7 Wikipedia 2009). HL7 Version 2 has conveyed "syntactic" interoperability among these vendor-based systems, such that data can be physically imported from one HL7 compliant system to another. The next advance, HL7 Version 3, not only provides syntactic interoperability, but also provides, very importantly for research usage, "semantic" interoperability. Although adoption of this HL7 version has been relatively slow by vendors and others, once in place it will allow for meaningful standardized understanding and interpretation of the data being exchanged across data systems.

Additionally standard information models for clinical data and research are being developed at this time as well. The Clinical Data Interchange Standards Consortium (CDISC) is a voluntary initiative to develop standards for clinical data across the Food and Drug Administration (FDA), pharmaceutical companies, and research institutions, ideally worldwide (Clinical Data Interchange Standards Consortium 2009; Clinical Data Interchange Standards Consortium Wikipedia 2009). The Biomedical Research Integrated Group (BRIDG) model is collaboration among HL7, CDISC, and the National Cancer Institute (NCI) to provide a common integrated data model for clinical research (Biomedical Research Integrated Domain Group 2009). These standard-setting initiatives, some of which are described in Chap. 9, will greatly enhance and support the ability to integrate EMR and CDMS data for research in the future.

2.6 Benefits of Integrating CDMS and EMR Systems

The integration of electronic records arising from the EMR and the CDMS could facilitate new interfaces between care and research environments, leading to great improvements in the scope and efficiency of research (Powell and Buchan 2005). Clinical narrative information, captured electronically as structured data or as transcribed "free text," when combined with other existing data, can dramatically increase the breadth and depth of information available for nonclinical applications (Safran et al. 2007).

Clinical trials, outcomes research, survival analyses, survey studies, and epidemiological research in cancer could all benefit from secondary use of EMR data for research purposes. Secondary uses of health data can expand knowledge about cancer diagnoses and treatments, strengthen understanding of healthcare systems' effectiveness and efficiency, support public health and security goals, and aid businesses in meeting customers' needs (Safran et al. 2007). Possible research benefits range from systematically generating hypotheses for research to eventually undertaking entire studies based only on electronic record data. Information for planning studies, such as prevalence and variance of conditions in local contexts, could be collected with relative ease (Powell and Buchan 2005).

Researchers can utilize secondary data to supplement their own data, to expand on or check the findings of the original studies, to test hypotheses or analyze relationships quite different from those analyzed and reported in the original study (Brown and Semradek 1992). Using longitudinal patient care data, they may discover or identify trends in relation to changes in the social and physical environment (Brown and Semradek 1992). Another evolving use of patient records data is to support clinical practice for the development of guidelines for clinical practice (Anonymous 1993). Such usage of EMR data also facilitates outcomes research, in which guideline performance and success of patient care can be evaluated and correlated, much as is being carried out within the National Comprehensive Cancer Network (NCCN) outcomes research project (Niland 1998).

Vital statistics are essential for determining the health needs of the population and for program planning and evaluation. Disease-specific mortality rates help pinpoint the major health problems of the population and target at-risk groups for interventions, and natality and infant mortality data help in planning maternal and child health programs (Brown and Semradek 1992). The crucial survival analyses required for such research can be greatly facilitated through the mortality data available through the EMR. In addition to utilizing information available through the EMR, national registers of diseases and treatments could be established more easily and economically with a coherent approach to security across agencies (Robertson 2003). This process could accelerate and expand epidemiological research, via disease registries encompassing well-characterized populations (Robertson 2003).

In the course of providing cancer care, practitioners with access to an EMR rely on this system to monitor patient progress, provide continuity of care, maintain patient care standards, and monitor quality of care. Another major benefit of secondary usage of clinical care data within research is that automated decision support could be incorporated into the conduct of interventional research studies to help ensure the safety of patients as they are being treated with highly experimental drugs. As an example, City of Hope Cancer Center has developed and incorporated into their monitoring of cancer clinical trials a system called the Cancer Automated Lab-based Adverse Event Grading Service (CALAEGS). The CALAEGS is fed laboratory results and normal ranges for clinical trial patients from the City of Hope EMR to provide automated grading of lab-based adverse events (AEs). The CALAEGS system has been proven to greatly improve the accuracy and completeness of AE reporting for the many thousands of lab tests that must be assessed for a given trial, compared with the former manual method (Niland et al. 2007).

2.7 Challenges of Integrating CDMS and EMR Systems

Rapidly evolving nationwide efforts for more widespread health information exchange must include work to address pressing issues of secondary health data usage (Safran et al. 2007). However, there are many challenges associated with achieving this complex and difficult goal. Secondary use of health data poses technical,

strategic, policy, process, and economic concerns related to the ability to collect, store, aggregate, link, and transmit health data broadly and repeatedly for legitimate purposes (Safran et al. 2007). The current lack of coherent policies and standard "good practices" for secondary use of health data impedes efforts to transform the US healthcare system (Safran et al. 2007). As new record systems are designed, records and record-keeping habits need to be studied to improve our processes and to identify redundancies that can be eliminated in the future (Institutes of Medicine 1991). Extreme care must be taken and failsafe processes put in place to ensure that the appropriate record linkage is occurring both across the various EMR systems that may contain data on the same patient and between the EMR data and the clinical research data. Some of the critical factors for meeting the challenges of EMR-clinical research system integration are described here.

2.7.1 Metadata Management

Metadata or "data about the data" are critical to successfully document, interpret, and analyze patient care or clinical research data. Two general forms of metadata exist, the "technical metadata" utilized by the programming staff and database architects to define the structure of the database, including the field types, lengths, table storage locations, etc. The technical metadata generally arise from the creation of the database itself and are therefore readily available and accessible from the database management system.

The other form is the "business metadata," including the data definitions, directives for collection, allowable code lists, creation date, sunset date, etc. The business metadata is critical from the data user's perspective, but is not so readily available, as it takes a major human manual curation effort to diligently create and maintain the business metadata for any given electronic data system. Tools for business metadata management are not widely accepted and standardized, and it is tempting and all too easy to create a database system and fail to document this critical information in a timely manner or at all. Best practices would dictate that the database elements cannot be created, changed, or deleted without requiring the attendant business metadata to be documented. Only through such documented information can the integrated EMR and CDMS information be valid or meaningful as it is analyzed and reported.

2.7.2 Data Quality Assurance

Whether data are entered into an EMR or CDMS, or integrated via a data warehouse, data quality checking is a mandatory process to ensure valid, accurate, complete data, particularly as in most cases the data entered into these systems are several steps removed from the original source of the information. In the case of

interventional research such as clinical trials, the original source of the data includes the caregiver generating the observations on patients, or the laboratory, blood bank, or other healthcare application that processes a patient's sample results, or at times the patients themselves, for example, via completion of home diaries. In observational research, the data may be provided directly by the patient, for example, surveys, but still could contain inaccuracies or be incomplete due to recall issues, or misunderstanding of his/her medical condition. The data also could arise from a secondary source once removed from the primary subject, such as a family member or caregiver, who may not have full accurate knowledge of the desired information. While billing and financial information may be quite useful for research purposes, the quantitative data of administrative records often are imprecise and unreliable (Brown and Semradek 1992).

Data quality assurance is a laborious and imperfect process. When data entry is involved in capturing the data within a CDMS, a traditional but time-consuming method to decrease data entry errors is the process of double data entry. This process may be carried out by the same person who initially keyed in the data or preferably by a second independent party. Once data have been screened for typographical errors, the entries can be further validated to check for logical errors, such as mistakenly entering the patient's year of birth as the current year. In addition, process errors may be detected, for example through a check of the subject's age to ensure that they are within the inclusion criteria for the study. These instances are flagged for review to determine if there is an error in the data, an incorrect process has occurred within the study conduct, or further medical clarification from the investigator or caregiver is required.

2.7.3 Data Completeness

To achieve linkages and the ability to aggregate data, several conditions must be met. A set of core data elements will need to be defined and recorded for all patient records, ideally including problem lists with current status and clinical rationale, as well as standard data within future patient records that can be drawn upon for research (Institute of Medicine 1991).

One investigation found that many items of information that a researcher might desire frequently are not available. For example, while sex and age were routinely noted in over 90% of cases, other basic demographic information was less frequently available: marital status in 79% of cases, race 40%, occupation 40%, religion 36%, and education 35% (Brown and Semradek 1992). The absence of such core data elements clearly will handicap certain research, such as efforts to relate illness to environmental factors. Clinicians have recognized that data collection is more accurate and complete when accomplished while the patient is still in the hospital, rather than through retrospective chart review, as missing elements could be obtained from physicians and definitions could be more consistently applied (Robertson 2003). Because data can be reviewed on a daily basis, omissions or

errors can be identified and corrected while the patient and their records are still immediately available (Robertson 2003).

Those who elect to use secondary data, whether researchers, practitioners, educators, administrators, or policy makers, have the obligation to evaluate the data they employ and to demand high quality and completeness. Otherwise, based on unsound data, research will be compromised and end, if not in failure, in less than optimal success (Brown and Semradek 1992).

2.7.4 Data Coding and Granularity

Coding of data is a critical process for the capability of generating analyzable information (Rangachari 2007). Two key areas that are not widely available in coded manner in the EMR, but are required within the CDMS are adverse event terms and medication names. In cancer the Common Terminology Criteria for Adverse Events (CTCAE) is the most common grading scale, and standard dictionaries of these terms can be loaded into the CDMS. Then the data items containing the adverse event terms or medication names can be linked to one of these dictionaries. An emerging standardized coding system for drugs is the RxNorm system (NLM 2009). Some systems allow for the storage of synonyms to allow the system to match common abbreviations and map them to the correct term. As an example, ASA could be mapped to Aspirin, a common notation.

Because every medical practice has distinct requirements, EMR systems usually need to be custom tailored (Electronic Medical Record Wikipedia 2009). The majority of EMR systems are based on templates that are initially general in scope. These templates can then be customized in cooperation with the system developer to better fit data entry based on a medical specialty, environment, or other specified needs. These templates tend to be customized individually by each organization, with few reusable standards in place. There are also EMR systems available that do not use templates for data entry and therefore can be easily personalized by each individual user. While this is advantageous in terms of flexibility for individualized patient care, the process leads to silos of information and lack of standardized information that can be shared across data systems and integrated with the CDMS. Further, secondary data often are aggregated to a less granular level, and this fact, or the unit by which data are aggregated, may render the information unusable for research purposes (Brown and Semradek 1992).

Risk adjustment is required not only to account for differences in patient characteristics across hospitals to enable comparison of hospitals' outcomes (such as mortality rates or the complication rates), but also to adjust risks within research analyses (Iezzoni 1997). Hospital coding accuracy is critical for ensuring accurate risk adjustment and, correspondingly, reliable comparative quality ratings (Rangachari 2007). Existing studies on hospital coding accuracy have viewed coding from a purely reimbursement perspective rather than a quality-measurement perspective or for research purposes (Rangachari 2007).

2.7.5 Data Access and Security

Secure management of electronic records from either the CDMS or EMR is a major concern to protect the confidentiality of the individuals involved. Such concerns are magnified further with regard to the potential privacy risk additionally posed by integrating information across the CDMS and the EMR. There is a potential lack of protection of PHI when used by entities not explicitly covered by HIPAA legislation or regulations (Safran et al. 2007). While providing a reasonable solution to this problem is not difficult, providing a perfect solution to the problem currently is impossible (Hammond et al. 1997). Patients must be reassured that no personally identifiable information will be used for research without the consent of the individual (Robertson 2003). Establishing role-based security can help achieve protection of the information by restricting access to particular types of information within the system based on the individual's need to access the data and then providing access only to the necessary types of data (Niland et al. 2006).

2.8 Conclusions

It can be seen that there are many advantages to secondary use of healthcare data for the purposes of clinical and translational research. Many different forms of cancer research can benefit from the integration of the EMR with the CDMS (Niland and Rouse 2006). Observational studies and case series may be conducted more rapidly, and new hypotheses generated through data mining. In epidemiological research, previously undetected patterns of response or toxicity could be detected more readily if a core set of uniform high-quality data were available for all patients. Clinical trials could be greatly expedited by using the EMR data to screen for potentially eligible subjects and to document their presenting characteristics if they enter into the trial. During the trial conduct, test results could be imported electronically from the EMR, so that automated decision support could help guard the safety on patients receiving highly experimental treatment. Outcomes research analyses could be facilitated by the availability of coded data on subjects' past history, comorbidity, treatments, and long-term outcomes.

However, there are also many challenges to achieving the full benefits of integrated data across the CDMS and the EMR. Quality, consistency, and standardized coding of the EMR data must be in place both within an institution and among institutions. Care must be taken to fully safeguard the integrated data, as computerized databases of personally identifiable information may be accessed, changed, or deleted more easily and by more people than with paper-based records. Metadata that carefully documents the definitions, conditions under which data arise, coding schemas available, etc. must be complete and readily available to the users of the integrated information.

Yet there is no doubt that the emerging EMR holds great promise for speeding biomedical discoveries through integration with the CDMS data. It is hoped that EMR adoption and standardization will proceed rapidly throughout the USA, and other countries worldwide, so that this promise can be realized.

References

Accredited Standards Committee (ASC) X12. Retrieved 17 August 2009. http://www.x12.org/

Adelman S, Moss LT (2000) Data warehouse project management. Addison-Wesley, Upper Saddle River, NJ

American National Standards Institute Accredited Standards Committee X12 (ANS ASC X12). In: Wikipedia, the free encyclopedia. Retrieved 12 August 2009 from http://en.wikipedia.org/wiki/ASC_X12

Anonymous (1993) Users and uses of patient records. Report of the Council on Scientific Affairs, American Medical Association. Arch Fam Med 2(6):678–681

Automated Decision Support Systems (ADSS). In: Wikipedia, the free encyclopedia. Retrieved August 4, 2009 from http://en.wikipedia.org/wiki/Automated_decision_support

Biomedical Research Integrated Domain Group (BRIDG). Retrieved 17 August 2009. http://bridgmodel.org/

Brown JS, Semradek J (1992) Secondary data on health-related subjects: major sources, uses, and limitations. Public Health Nurs 9(3):162–171

CDC National Center for Health Statistics (2006) More physicians using medical records. http://www.cdc.gov/media/pressrel/a060721.htm?s_cid=mediarel_a060721. Accessed 20 August 2009

Center for Medicare and Medicaid Services (CMS) Fact Sheet (2009) Medicare and medicaid health information technology: Title IV of the American recovery and reinvestment act. Retrieved 12 August 2009. http://www.cms.hhs.gov/apps/media/fact_sheets.asp

Certification Commission for Health Information Technology (CCHIT). Retrieved 17 August 2009. http://www.cchit.org/

Certification Commission for Healthcare Information Technology (CCHIT). In: Wikipedia, the free encyclopedia. Retrieved August 4, 2009 from http://en.wikipedia.org/wiki/Certification_Commission_for_Healthcare_Information_Technology

Choi B, Drozdetski S, Hackett M, Lu C, Rottenberg C, Yu L, Hunscher D, Clauw D (2005) Usability comparison of three clinical trial management systems. AMIA Annu Symp Proc 2005:921

Clinical Data Interchange Standards Consortium (CDISC) (2009) Retrieved 15 August 2009. http://www.cdisc.org/

Clinical Data Management System (CDMS). In: Wikipedia, the free encyclopedia. Retrieved 4 August 2009 from http://en.wikipedia.org/wiki/Clinical_data_management_system

Digital Imaging and Communications in Medicine. In: Wikipedia, the free encyclopedia. Retrieved 12 August 2009 from http://en.wikipedia.org/wiki/Digital_Imaging_and_Communications_in_Medicine

Electronic Data Capture (EDC). In: Wikipedia, the free encyclopedia. Retrieved 4 August 2009 from http://en.wikipedia.org/wiki/Electronic_data_capture

Electronic Health Record (EHR). In; Wikipedia, the free encyclopedia. Retrieved 4 August 2009 from http://en.wikipedia.org/wiki/Electronic_health_record

Electronic Medical Record (EMR). Electronic Data Capture (EDC). In: Wikipedia, the free encyclopedia. Retrieved 4 August 2009 from http://en.wikipedia.org/wiki/Electronic_medical_record

Gray P, Watson HJ (1998) Decision support in data warehouse. Prentice Hall, Upper Saddle River, NJ

Greenes RA, Pappalardo AN, Marble CW, Barnett GO (1969) Design and implementation of a clinical data management system. Comput Biomed Res 2(5):469–485

Hammond WE, Hales JW, Lobach DF, Straube MJ (1997) Integration of a computer-based patient record system into the primary care setting. Comput Nurs 15(2 Supp 1):S61–S68

Handleman D (2005) Electronic data capture: when will it replace paper? Retrieved 4 August 2009. http://www.sas.com/news/feature/hls/sep05edc.html

Health Information Privacy. In: U.S. Department of Health & Human Services. Retrieved 12 August 2009. http://www.hhs.gov/ocr/privacy/

Health Information Technology for Economic and Clinical Health (HITECH) Act (2009) H.R.1 – American Recovery and Reinvestment Act of 2009. Retrieved 20 August 2009. http://www.opencongress.org/bill/111-h1/show

Health Insurance Portability and Accountability Act. In: Wikipedia, the free encyclopedia. Retrieved 12 August 2009 from http://en.wikipedia.org/wiki/HIPAA

Health Level Seven (HL7). Retrieved 17 August 2009. http://www.hl7.org/

Health Level Seven (HL7). In: Wikipedia, the free encyclopedia. Retrieved 15 August 2009 from http://en.wikipedia.org/wiki/Health_Level_7

Iezzoni L (1997) Risk adjustment for measuring health care outcomes. Health Administration Press, Chicago, IL

Institute of Medicine (1991) The computer-based patient record: an essential technology for health care. National Academy Press, Washington, DC

Medicare Part B Imaging Services (2008) United States government accountability office report to congressional requesters. http://www.gao.gov/new.items/d08452.pdf. Accessed 20 August 2009

National Archives and Records Administration (2008) Long-term usability of optical media. Retrieved 20 August 2009. http://206.180.235.135/bytopic/electronic-records/electronic-storage-media/critiss.html

National Cancer Institute. Electronic Medical Record, Dictionary of Cancer Terms. Retrieved 19 August 2009. http://www.cancer.gov/Templates/db_alpha.aspx?CdrID=561399

National Center for Health Statistics (2006) Electronic medical record use by office-based physicians, United States, 2005. http://www.cdc.gov/nchs/products/pubs/pubd/hestats/electronic/electronic.htm. Accessed 20 August 2009

National Library of Medicine (2009) National Institutes of Health, Unified Medical Language System, RxNorm. Retrieved 19 August 2009. http://www.nlm.nih.gov/research/umls/rxnorm/

Niland JC (1998) NCCN internet-based data system for the conduct of outcomes research. Oncology 12:11

Niland J, Rouse L (2006) Clinical research needs. In: Lehmann H, Abbott P, Roderer N et al (eds) Aspects of electronic health record-system, 2nd edn. Springer, New York

Niland JC, Rouse L, Stahl DC (2006) An informatics blueprint for healthcare quality. J Am Med Assoc 13(4):402–417

Niland JC, Pannoni S, Neat J, Sarbora R, Lee J (2007). Cancer Automated Lab Adverse Event Grading Service (CALAEGS). American Medical Informatics Association 'Clinical Research Informatics EXPO' Posters and Demonstrations

Payne PR, Greaves AW, Kipps TJ (2003) CRC Clinical Trials Management System (CTMS): an integrated information management solution for collaborative clinical research. AMIA Annu Symp Proc 2003:967

Powell J, Buchan I (2005) Electronic health records should support clinical research. J Am Med Internet Res 7(1):e4

Rangachari P (2007) Coding for quality measurement: the relationship between hospital structural characteristics and coding accuracy from the perspective of quality measurement. Perspect Health Inf Manag 4:3

Robertson J (2003) Cardiovascular point of care initiative: enhancements in clinical data management. Qual Manag Health Care 12(2):115–121

Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE (2007) Toward a national framework for the secondary use of health data: An American informatics association white paper. J Am Med Assoc 14:1–9

Stam H, van Ginneken AM (1995) Computer-based patient record with a cardiologic extension. Medinfo 8(Pt 2):1666

Summerhayes S (2002) CDM regulations procedures manual. Blackwell, London

Tai BC, Seldrup J (2000) A review of software for data management, design and analysis of clinical trials. Ann Acad Med Singapore 29(5):576–581

Thiru K, Hassey A, Sullivan F (2003) Systematic review of scope and quality of electronic patient record data in primary care. Br Med J 326:1070

Turban E, Leidner D, McLean E, Wetherbe J (1997) Information technology for management: transforming organizations in the digital economy, 6th edn. Wiley, Danvers, MA

Winkelman WJ, Leonard KJ (2004) Overcoming structure constraints to patient utilization of electronic medical records: a critical review and proposal for an evaluation framework. J Am Med Inform Assoc 11:151–161