

## **Reflections on preparing electronic health record data for clinical research**

Suzanne L. West<sup>1,2</sup>, Catherine Blake<sup>3,4</sup>, Zhiwen Liu<sup>2</sup>, J. Nikki McKoy<sup>5</sup>, Maryann D.

Oertel<sup>6,7</sup>, Timothy S. Carey<sup>8,9</sup>

1. Health, Social and Economics Research, RTI International, Research Triangle Park, North Carolina 27709-2194
2. Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7435
3. School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360
4. UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7295
5. Institute for Medicine and Public Health, Vanderbilt University, Nashville, TN 37203-1738
6. Drug Information Services, Department of Pharmacy, UNC Hospitals, Chapel Hill, NC 27599-7600
7. Assistant Clinical Professor, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
8. Cecil G. Sheps Center for Health Services Research, University of North Carolina, Chapel Hill, NC 27599-7590
9. Department of Medicine, University of North Carolina, Chapel Hill, NC 27599-4228

Correspondence to: Suzanne L. West, MPH, PhD, RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194. Email: [swest@rti.org](mailto:swest@rti.org); fax: 919-541-7384

## **Abstract**

The adoption of electronic health records (EHRs) offers the potential to improve the delivery, quality, and continuity of clinical care, but widespread use has not yet occurred. In this paper, we describe a process for using clinical (production) data that were derived from outpatient and inpatient visits at a university teaching hospital for clinical research, a use for which the data and its structure were not originally designed. Similar data exist at many outpatient and inpatient clinical facilities, and we believe that our insights are relevant to electronically captured medical data regardless of its origin. We describe the approaches taken to ensure compliance with the Health Insurance Portability and Accountability Act (HIPAA) and to leverage the vast stores of structured and unstructured data that are currently underused. We conclude by reflecting on what we would have done differently and by making recommendations to streamline the process.

## **Keywords**

electronic health record, medical records systems, computerized; confidentiality; HIPAA

# 1 Introduction

Publications since the 1960s have suggested that computerized medical information systems, especially electronic health records (EHR), can improve the quality and efficiency of patient care[1-5]. Nearly 50 years later, we still do not have a common structure and content for an EHR. A report to the Chairman on Health Information Technology found that “There is a lack of consensus on what constitutes an EHR, and thus multiple definitions and names exist for EHRs, depending on the functions included. An EHR generally includes (1) a longitudinal collection of electronic health information about the health of an individual or the care provided, (2) immediate electronic access to patient- and population-level information by authorized users, (3) decision support to enhance the quality, safety, and efficiency of patient care, and (4) support of efficient processes for health care delivery” [6](page 11) Today, more than 50 different commercial EHR systems are available [7], but like the variation in the composition of an EHR, coding procedures and software differ and have not been standardized . Given the diversity of information needs in healthcare settings, any one EHR system likely cannot meet the full range of requirements that various healthcare providers and administrators desire [8].

The complexity and longitudinal requirements for EHR presents data challenges that may not be immediately apparent. Health care data originates from both the outpatient and inpatient setting and comprises both numeric and textual information for medications diagnostic tests, demographics, problem lists, staff notes; images for radiology; and knowledge bases for clinical guidelines and reminders.[8-12] Moreover some of the EHR data that, in its original format, would be textual in nature such as problem lists and laboratory results, are categorized and therefore structured, whereas physician notes and

radiology or pathology results are unstructured (free) text. This may limit the usability of these data for research. Further, many practices choose not to implement all of the EHR's capabilities, resulting in disparate categories of data collected.[13, 14] These issues do not even address the fact that these data may be entered and managed by different individuals within the same institution which adds another dimension to their variability.

Ideally, the information organized in an EHR will enhance the quality of clinical care and facilitate research. [7–10] However the EHR is a necessary, but not sufficient condition to improve quality care. Specifically clinicians must know what constitutes 'quality' for the given medical condition, thus evidence-based guidelines also play a role in improving the quality of care[15], as does adherence to such guidelines. In a recent study, the quality of care for patients with heart failure [16] was not enhanced when guidelines were installed within an EHR. In another study comparing two similar primary care practices, one with and one without an EHR, glycosylated hemoglobin (HbA1c) and lipid tests were ordered more frequently in the practice using the EHR, but this did not translate to better glycemic control[17]. Inclusion of evidence-based performance indicators into EHRs will likely improve the quality of care measurably [18].

We sought to use clinical information from an EHR to determine whether these data could be used to assess the quality of care for patients with newly diagnosed diabetes mellitus type 2 (T2DM) with regard to medication treatment patterns. The clinical information system we evaluated had not been used for research previously. We describe the processes used for de-identification and preprocessing of the data, manual review to understand the available data, algorithms for identifying patients with newly diagnosed T2DM, and use of text-mining methods for retrieval of medication data.

We use a case study approach to describe the types of clinical data that can be used to address quality of care and to outline the processes we used to tap into existing records. Our medical setting is a large academic hospital, which we describe in section 2. Section 3 provides a developmental history of the “homegrown” EHR used at the academic medical center. We then describe the process we used to collect and prepare data for subsequent experiments in section 4. Section 5 reflects on our experiences using an EHR developed for clinical use as a research tool.

## **2 Medical Setting**

This research used data from the University of North Carolina’s Health Care System (UNCHCS). UNC Hospitals, affiliated with UNC Chapel Hill School of Medicine, comprises 4 hospitals housing >700 beds served by ~1,000 attending and 550 resident physicians. In 2002, there were ~700,000 clinic visits, 30,000 non-obstetric discharges, and 3,000 deliveries. All hospital facilities, patients, and staff are served by one clinical information system (WebCIS).

## **3 Evolution of WebCIS**

Efforts to develop an EHR began in the early 1980s, with the overarching goal of improving quality of care and education in UNCHCS’s general medicine practice. Because outpatient records were unavailable for 10%–15% of patient visits, an outpatient ‘Mini-Medical Record’ (MMR) was developed that contained essential information for practicing clinicians[19]: patient demographics linked to the registration system, a problem list, vital signs, a medication list that doubled as a prescription document for internal use, and limited health maintenance prompts.

During the mid- to late 1980s, the system was disseminated for use in essentially all outpatient practices at UNC. Although the MMR proved to be moderately useful for outpatient care, it had four main limitations. First, the interval history, physical examination, plan, laboratory data, and hospital discharge summaries were unavailable. Second, medication information was inadequately coded and internal prescriptions were accepted only by the UNCHCS pharmacy, leading to both duplicate and missing medication data. Third, there were substantial personnel costs for data entry by clerks of the information written by clinicians. Finally, the system had been written in FORTRAN and was becoming increasingly difficult to maintain.

In 1990, the UNCHCS Information Services Division (ISD) and medical staff began to develop a fully computerized replacement for the MMR. The initial version of this system, the Clinical Information System (CIS) version 1.0, went live in 1991. It included the elements from the MMR plus access to discharge summaries and radiology reports. Version 2.0 (early 1993) included information from the inpatient wards. Version 2.1 allowed access to real-time laboratory test information. Finally, a security system was incorporated that recorded the date and time of each transaction and used a single user login for essentially all clinical data. This system was used until the Web-based version (WebCIS) was developed.

Several principles guided the development of WebCIS. First was to facilitate access to current and previous records—including problem lists, medication lists, laboratory data, clinical information, and reports with the same ‘look’ and navigation rules—throughout UNCHCS using a secure, Web-based system. The system also had to allow patients to be followed in inpatient and outpatient settings. It was designed to link with legacy systems

such as patient registration, provider lists, and allergy systems, in real time and bidirectionally to the degree possible. Prompts were to be added gradually, as were links with vendor systems such as computerized physician order entry. Standard coding, such as International Classification of Disease (ICD-9-CM) codes for diagnoses and National Drug Codes (NDC) for drugs, was implemented when possible rather than having free text data entry to facilitate research and analyses. A proprietary relational database, IBM's DB2, was to be used for data storage and production.

WebCIS became the primary clinical record system used at UNCHCS in April 2001. Further refinements have been made over time. For example, in October of 2002, WebCIS 1.5 delivered telephone message services and linkage to the Picture Archival System (PACS) for online display of diagnostic radiography and imaging scans. Version 2.0 fully replaced the outpatient MMR in late 2004 and incorporated direct entry of coded drugs, allergies, and problem lists with the ability to keep a dated, annotated record of active/inactive entries; the ability to print new and refill prescriptions; direct entry of coded vital signs and nurse's notes for outpatient areas; automatic alerts to providers for health maintenance, immunizations, disease management, completed ancillary tests for current outpatients, inpatient admissions, and inpatient deaths; expansion of Personal and Group patient lists to allow multiple personal lists and the ability to create or join new group lists; and electronic signature of clinical notes for Medicare patients.

Version 2.5, released in April 2005, added a forms tool allowing direct data entry of inpatient notes (history, physical examination, progress notes, consults, procedure notes, operative notes) and direct transmission of e-prescriptions to area pharmacies. This tool pulls the relevant data from problems, medications, allergies, and the history into the note,

thus eliminating double entry and streamlining documentation. The prescription-writing capability increased physician satisfaction and use of WebCIS dramatically.

The current WebCIS version includes information from the clinical and administrative areas shown in Table 1. The DB2 database consists of 47 tables (table names and content are available upon request). The medical record number (MRNO) is the primary key for all tables, and the patient name is the unofficial secondary key.

## **4 Preparing data for Clinical Research**

The goal of the clinical research case study was to assess medication patterns, a quality indicator for diabetes treatment. We focused on patients with newly diagnosed T2DM who often have comorbid conditions such as hypertension and dyslipidemia. We first performed a pilot study using data from five such patients, followed by the full case study.

### ***4.1 Step 1: Data Collection***

The main data files used were the Patient Problems file, which reflects diagnoses of current and chronic conditions (TPRBPAT.txt); the Medication Prescribed file (TFDBPAT.txt); the Laboratory file (TLABRSSC.txt); and the Visit Transcription file (TTRNTEXT.txt) (Figure 1). Two variables in the TPRBPAT.txt file—CPK\_ICD9\_CODE, which contains the ICD-9-CM code clinicians use to record the reason for a visit, and C\_ONSET\_DATE, which corresponds with the date of entry for the ICD-9-CM code—appeared particularly important for our research. Although these variables seemed to indicate when a condition was first diagnosed, this assumption was true only for acute conditions. For chronic conditions such as diabetes, clinicians entered a new date each time they provided care, although the file also retained older dates.

Given that the focus of our project was to assess treatment patterns for patients with

newly diagnosed T2DM, the validity of the medication data was critical. The TFDBPAT.txt file contains the name, dosing, regimen, and dates of use for prescribed drugs, and the inactivation date refers to the date on which a new prescription is written for the same regimen of the same drug (Figure 1). If the clinician prescribes a drug only once with no refills (such as with antibiotics), the inactivation date is typically blank, preventing determination of the duration of drug use. Further, medication information was inadequately recorded in WebCIS before 2002. Even after implementation of e-prescription transmissions to pharmacies in that year, data completeness remained a concern. We discuss strategies used to address this issue below.

The Laboratory file (TLABRSSC.txt) was particularly useful, as it contained lab results from as far back as January 1994 (Figure 1). The UNCHCS laboratories use their own coding system [20], not the industry standard [21], which had been developed 2 years after the earliest UNC laboratory results were available. We faced consistency issues using the WebCIS laboratory data because the codes used varied depending on where the patient's sample was taken. For example, 8 different codes were used to identify HbA1c level within the WebCIS laboratory files.

The Visit Transcription file (TTRNTEXT.txt) reporting outpatient care was one of the most informative files in WebCIS, but text mining was required to realize its potential for research. The entries describe family and social history; use of alcohol, tobacco, and illicit drugs; medication information; historical data regarding disease onset; and clinical quality indicators, which for diabetes includes contraindications to sulfonylureas, metformin, angiotensin-converting enzyme inhibitors, or angiotensin receptor blockers; foot examinations; and referrals for ocular examinations (Figure 2).

## 4.2 Step 2: Data preprocessing

As these data were from a “live” EHR, they had to be downloaded from the mainframe system in a format usable by researchers. UNCHCS ISD staff worked with us to provide data extracts as text files in which the majority of data were de-identified (the unstructured data were not); the first download of data also contained extraneous characters. We describe our processes for handling these issues below.

Data conversion. ISD staff loaded the data extracts with pre-specified data structure into project-specific, password-protected regions on a secure server that was backed up daily. The first data extract contained null characters created when data from numeric fields containing integers had been converted to character fields. ISD programmers minimized this problem in later data extracts, but some extraneous characters remained. Our project programmers developed a program to identify and delete these extraneous characters from files so the data could be more easily uploaded into SAS (Cary, NC) .

De-identification. With passage of the Health Insurance Portability and Accountability Act (HIPAA) in 1996, which went into effect on April 14, 2003[22], a Privacy Rule was enacted to protect patient privacy[23]. The Rule applies to ‘covered entities,’ comprising those who generate data (e.g., medical groups and physicians) and those who manage or handle personally identifiable health information (e.g., payers). To comply with this rule, we obtained a limited dataset<sup>1</sup> from UNCHCS, the covered entity.

De-identification of structured and unstructured data required different strategies. De-identifying structured data was straightforward. Attributes with identifying information

---

<sup>1</sup>Limited Data Set under 45 CFR §164.514(e): ‘Under certain circumstances, a covered entity may use and disclose protected health information (PHI) in a limited data set for research, public health, and health care operations purposes. The privacy regulation identifies a list of identifiers that must be removed from data in order for it to be considered a ‘limited data set.’ Once removed, the information is not de-identified – it is still PHI governed by the privacy regulation. A data use agreement must be signed by those wishing to use limited datasets.’

such as 'name,' 'phone number,' or 'Social Security number (SSN)' were simply omitted from the data extraction. We encrypted other necessary fields such as MRNO as required. Fortunately, most WebCIS data used in this project consisted of structured data, which allowed us to remove direct identifiers easily and accurately.

We required a more sophisticated process to de-identify data in text files, such as the transcriptions of patient visits. UNC clinicians use voice-recognition systems from external vendors to dictate visit transcriptions. The resulting transcription data file contains long text fields of largely unstructured data, leading to variability in headers, formatting, etc. A single field in one record could contain a series of text strings formatted in a particular way, whereas the same field in another record could contain different types of text data formatted completely differently. Further, the same field in a third record could contain a text translation of a voice memo. Some specialty clinics also use other systems to collect information during care and then 'push' the selected data into WebCIS. Speech-recognition software also is used to convert dictated memos to text for inclusion in the record. These approaches present problems when trying to analyze individual elements within the chunks.

Finding data containing patient identifiers for removal also was difficult. Some records contained attribute names that could be used as markers for identifying data, such as 'Name:', 'Patient:', or 'MRNO:'. For some records, a text-filtering program searched for these markers and replaced the text immediately following. However, markers were inconsistent across specialty clinic records, some records contained identifiers but no markers, and the number of characters to be replaced after the markers varied. The body of the note also could contain a patient's name or MRNO anywhere within a 2000-character

text field. Moreover, physicians refer to patients in varied ways: ‘Ms. Doe,’ ‘Jane,’ ‘Jane Doe,’ etc. Although ‘Jane’ in one sentence may be insufficient to identify an individual, the combination of ‘Jane’ with ‘Ms. Doe’ in a later sentence would be. An effective de-identification program must be flexible enough to handle these variations and others that might emerge. Given that one extracted file can contain 2 million lines of text, however, it was not practical to explore a file of this size manually looking for all possible variations.

We therefore considered other ways to de-identify extracted data. The ISD group generated a dataset containing direct identifiers such as first name, last name, MRNO, address, phone number, SSN, etc. for each patient with records in the extracted transcription file. This ‘key’ dataset also contained a ‘fake medical record number’ (fake\_MRNo) that served as a linking field between the key dataset and the transcription file for each patient. Using the direct identifiers, we simplified the problem to replacing the identifiers in the research dataset without having to locate them first.

Because the extract was pulled directly from a live clinical information system, the extraction routines were required to run at night, fitting into the ‘mainframe’ queue with other nightly jobs. External programs, such as the scrubbing program, were not permitted to run in UNCHCS’ mainframe environment. In addition, to abide by HIPAA regulations and our data-use agreement, data scrubbing had to be done at the ISD site to prevent unauthorized disclosure of direct identifiers. Our solution was to write the scrubbing program in the Practical Extraction and Report Language (PERL) to run on a personal computer at the ISD offices. The open-source licensing of PERL allowed easy installation at no direct cost, and it provided the necessary file handling, regular expression, and data structure functions. We loaded the program onto a UNCHCS computer, deidentified the

unstructured data onsite, and then physically transported the file to our research offices.

The first step of the scrubbing program was to read all 12,910 records from the key dataset into a data structure for quick access. Each record contained a set of direct identifiers and a unique Fake\_MRNo for a specific patient. Figure 3 shows a hash table of associative arrays. The hash table stored these records in memory and provided fast access to individual records and the direct identifiers contained therein. The Fake\_MRNo was used to link research dataset records to key dataset records. After the program was run, the output data set had the word 'REMOVED' in place of the identifiable text.

The above process works only if there is a 1:1 relationship between the patient's name and MRNO. Using the example in Figure 1, if Jane Doe with SSN 234568795 has more than one MRNO, then only medical record 359785 will be de-identified; the other medical record number(s) will not.

### ***4.3 Step 3: Pilot Study***

We asked the ISD team to extract all available clinical data in de-identified form for 5 patients with diabetes. These data were downloaded from the 'live' system as 41 separate text files (the 5 test patients lacked data for 6 of the 47 tables). We reviewed these data manually, using the dates from the visits, laboratory tests, prescriptions, and diagnoses to understand the patients' clinical problems. The review revealed critical information about the completeness of the structured data and the value of the unstructured data.

### ***4.4 Step 4: Full Study***

One of the most challenging tasks in using electronic databases of U.S. patients (i.e., administrative claims or EHRs) is identifying the first diagnosis of a chronic condition. Each time patients switch health plans or healthcare providers, which often occurs in the

highly mobile U.S. society, the longitudinal nature of the data is lost. A further complication is referral patterns: As a tertiary-care, academic medical center, UNCHCS often sees patients who have been referred from their local primary care physicians. Thus patients seen at UNCHCS may be seen only sporadically, giving an incomplete picture of their care over time.

To address these potential limitations we developed inclusion and exclusion criteria that differentiated patients seen regularly and followed by UNCHCS practitioners from those seen only episodically for emergency or referral care (Figure 4). Patients seen at UNC several times a year and who underwent periodic HbA1c testing were using UNCHCS for their primary care. In contrast, those with sporadic visits and rare HbA1c tests were likely being followed outside our system. Thus, a primary component of the inclusion criteria was the number of HbA1c tests patients had over time.

We obtained a data extract for all patients who had an HbA1c test on or after January 1, 2002. The original research plan was to identify patients with a new diagnosis of diabetes as of January 1, 2003 and compare their cardiovascular outcomes according to the initial antidiabetic medications they had received. However, because many diabetic patients are often managed first by diet modification alone, we were concerned that a 2-year follow-up (2003–2005) might not provide sufficiently diverse treatment patterns to observe longitudinal trends in antidiabetic medications or clinical practice for assessing quality of care. We therefore revised the research plan to define the base population as patients who had a first HbA1c test recorded in WebCIS after January 1, 2001 and who had 2 or more HbA1c tests recorded in the laboratory file. We requested de-identified demographic, visit, medication, electrocardiography, patient problems, vital status notes, and transcription files

for 12,424 patients with multiple HbA1c tests (Figure 4).

Based on an extensive review of the laboratory, patient visit, and patient problem files for these 12,424 patients, we developed a set of exclusion and inclusion criteria for identifying patients with newly diagnosed T2DM who were regularly seen at UNC for their care (Table 2, Set 1). After applying these criteria, 2386 patients appeared to have newly diagnosed diabetes. We then excluded patients who received only insulin, because we could not easily distinguish between patients with type 1 versus type 2 diabetes in this group and could not track dosing with any confidence. After this exclusion, 1933 diabetic patients remained.

We wished to further examine the ability to identify patients with a new diagnosis of diabetes after January 1, 2001 who were not seen at UNCHCS as a referral or for emergency care. The goal was to develop an algorithm that would be highly specific for newly diagnosed diabetes. To that end, we mined the deidentified transcription data to capture all occurrences of three text strings: diabetes (diabet\*), insulin, and antidiabetic medication (using a list of all brand and generic names of antidiabetic drugs). The text data-mining procedure is described elsewhere [24].

We read the resulting text-processed file into SAS and identified 677 patients who had a mention of any of the three text strings before January 1, 2001. A visual review of these 677 records indicated many different reasons for capture of these patients, such as ‘mother has non-insulin-dependent diabetes,’ which would have qualified for two of the three text strings. Because SAS could not be used for this determination, each of the 677 records underwent dual independent review by the investigators using the criteria shown in Table 2 (Set 2). From this manual review, we excluded another 269 patients, leaving 1664 patients

presumed to be newly diagnosed with diabetes after January 1, 2001 (Figure 4).

## **5 Reflecting on EHRs for Clinical Research**

Our goal was to determine whether medication treatment information from an academic medical center EHR, WebCIS, could be used to evaluate the quality of care provided to diabetes patients. While WebCIS spans many subspecialties from both the inpatient and outpatient setting, an important factor that limited its utility for clinical and health services research was the incompleteness of the data, particularly the medication data. Some practitioners did not use the drug-entry fields in WebCIS until the system allowed them to print outpatient prescriptions and/or transmit them electronically to outside pharmacies. We addressed this challenge by collecting data from multiple sources including the transcribed notes from the patient visit which contained drug names, doses, and regimens which we captured via text mining. These sources were useful in establishing when a drug treatment strategy began, but neither the structured WebCIS data nor the transcription data was complete with respect to when the clinician discontinued the drug. We did not have access to pharmacy claims in addition to prescribing data for these patients; some have argued that such data more accurately reflect medical exposure because they indicate medications that are dispensed to the patient.[25]

Another challenge was the difficulty in discriminating new versus referred cases of diabetes. Because UNCHCS is a tertiary medical care system, some patients received their specialist care at UNCHCS but sought their generalist care from community practitioners. This pattern would give the appearance of “new-onset” diabetes when using WebCIS for research projects, but in reality, the diabetes had been present for some time but treated elsewhere. Thus, the fragmentation of the US healthcare system constitutes a major

obstacle for longitudinal research on the population, but also for providing and therefore, assessing the quality of patient care. Patients' use of multiple providers and changes in health insurer coverage limits researchers' and evaluators' ability to assess the adequacy of care over long periods of time. Similar issues occur with data from the US Veterans Administration because veterans may seek care both from the Veterans Administration facilities and community practitioners. In the United States, fragmentation of care is less of an issue for integrated healthcare systems such as staff model health maintenance organizations like the Kaiser Permanente Medical Groups.

The fragmentation of healthcare in the US occurs at all levels of care and payment: at the national, state, community, and practice levels.[26] For this project, fragmentation at the practice or provider level has made it difficult to determine when a patient was newly diagnosed with diabetes and how well he or she is being treated for this condition. The Commonwealth Report also says that community-based providers practice independently without considering that the patient is likely seeing other providers, which confirms what we noted using UNCHCS's WebCIS data. Using the information from only one provider's EHR does not allow a complete view of all of the patient's care, making it very difficult to discern whether the patient has received high-quality care. While the US system is particularly fraught with these issues, fragmentation occurs in other countries as well, including Finland [27] Israel and Canada [28 ] but to lesser extents.

Our experience with EHR data from this project makes us well poised to make recommendations about the characteristics of EHRs that would streamline their use in clinical research. The ability to assemble all clinical information for a patient, including diagnoses, medications, procedures, laboratory tests and results, and radiography

information, across all providers is critical for assessing care quality. We recommend a system that allows linkage across healthcare providers and insurers, either by assigning everyone an identifier, similar to what is done in the United Kingdom, Scandinavia, and other countries[29] or facilitate data linkage using a probability matching algorithm with a connection broker called a Record Locator Service—similar to a directory of providers and insurers for each patient.[30] Despite our recommendations, some Americans are against any type of linkage across healthcare data files.[31]

The second recommendation is to incorporate as many structured variable fields as possible in EHRs to reduce the need for extensive text mining, including the use of standard coding nomenclatures for entering disease, laboratory, and medication information. Encouraging providers to standardize entry procedures for recording care when instituting EHRs in clinical settings will aid in the use of these data research and assessing care quality. For example, if clinicians are instructed to record patient symptoms until a definitive diagnosis is made, then it is easier to determine the exact date on which a particular disease was diagnosed.

The third recommendation is to avoid the use of patient identifiers within the transcribed note. Phrases such as, “Mrs. Katherine Jones is a very pleasant 86 year old white female...” are used by providers to personalize the note. However, this information is not allowed according to the privacy rule and must be eliminated before the data leave the covered entity. It would be much easier to avoid including identifiable information in the first place.

## **6 Acknowledgements**

This project was funded under Contract No. HHS290-2005-0040-I from the Agency for

Healthcare Research and Quality, U.S. Department of Health and Human Services, as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program. The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

The authors thank Ms. Deborah Morris, Sergei Glazunov, and Raj Gopalan for assisting with the data extraction and for providing valuable input on the structure and history of WebCIS. We also thank Bob Schwartz, Frances Ochart, and Roger Akers for their expert programming and Drs. Mary Roth, Robert Malone, and Michael Pignone for their clinical input.

## 7 References

1. Barnett, G.O., *Computers in patient care*. N Engl J Med, 1968. **279**(24): p. 1321-7.
2. Schwartz, W.B., *Medicine and the computer. The promise and problems of change*. N Engl J Med, 1970. **283**(23): p. 1257-64.
3. Stead, W.W. and W.E. Hammond, *Computerized medical records. A new resource for clinical decision making*. J Med Syst, 1983. **7**(3): p. 213-20.
4. Barnett, G.O., *The application of computer-based medical-record systems in ambulatory practice*. N Engl J Med, 1984. **310**(25): p. 1643-50.
5. Hillestad, R., et al., *Can electronic medical record systems transform health care? Potential health benefits, savings, and costs*. Health Aff (Millwood), 2005. **24**(5): p. 1103-17.
6. Office, U.S.G.A., *HHS Is Taking Steps to Develop a National Strategy*, in *Health Information Technology*. 2005.
7. American Academy of Family Physicians. *Partners for Patients Electronic Health Record Market Survey*. [PDF] March 01, 2005 [cited; Available from: [www.centerforhit.org/PreBuilt/chit\\_2005p4pvendsurv.pdf](http://www.centerforhit.org/PreBuilt/chit_2005p4pvendsurv.pdf)]
8. Shortliffe, E.H. and J.J. Cimino, eds. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Health Information Series, ed. K.J. Hannah and M.J. Ball. 2006, Springer Science+Business Media: New York.
9. Institute of Medicine, *The Computer-based Patient Record : An Essential Technology for Health. Revised edition.*, ed. R.S. Dick, E.B. Steen, and D.E. Detmer. 1997, Washington, D.C. : National Academies Press. 234.

10. Lehmann, H.P., *Aspects of electronic health record systems*. 2nd ed. Health informatics. 2006, New York: Springer. xvii, 483 p.
11. Follen, M., et al., *Implementing health information technology to improve the process of health care delivery: a case study*. Dis Manag, 2007. **10**(4): p. 208-15.
12. Poon, E.G., et al., *Assessing the level of healthcare information technology adoption in the United States: a snapshot*. BMC Med Inform Decis Mak, 2006. **6**: p. 1.
13. Hing, E.S., C.W. Burt, and D.A. Woodwell, *Electronic medical record use by office-based physicians and their practices: United States, 2006*. Adv Data, 2007(393): p. 1-7.
14. Gans, D., et al., *Medical groups' adoption of electronic health records and information systems*. Health Aff (Millwood), 2005. **24**(5): p. 1323-33.
15. Garber, A.M., *Evidence-based guidelines as a foundation for performance incentives*. Health Aff (Millwood), 2005. **24**(1): p. 174-9.
16. Baker, D.W., et al., *Automated review of electronic health records to assess quality of care for outpatients with heart failure*. Ann Intern Med, 2007. **146**(4): p. 270-7.
17. O'Connor, P.J., et al., *Impact of an electronic medical record on diabetes quality of care*. Ann Fam Med, 2005. **3**(4): p. 300-6.
18. Feifer, C., et al., *Different paths to high-quality care: three archetypes of top-performing practice sites*. Ann Fam Med, 2007. **5**(3): p. 233-41.
19. Hammond, J.E., et al., *Making the transition from information systems of the 1970s to medical information systems of the 1990s: the role of the physician's workstation*. J Med Syst, 1991. **15**(3): p. 257-67.

20. Hammond, W.E., *The making and adoption of health data standards*. Health Aff (Millwood), 2005. **24**(5): p. 1205-13.
21. Regenstrief Institute Inc. *Logical Observation Identifiers Names and Codes (LOINC®)*. [accessed 2006 November 26]; Available from: <http://www.regenstrief.org/medinformatics/loinc/>.
22. US Department of Health & Human Services. *Protecting the Privacy of Patients' Health Information*. 2003 [accessed 2007 December 20]; Available from: <http://www.hhs.gov/news/facts/privacy.html>.
23. US Department of Health & Human Services. *Medical Privacy - National Standards to Protect the Privacy of Personal Health Information*. 2003 [accessed 2007 December 20]; Available from: <http://www.hhs.gov/ocr/hipaa/>.
24. Kraus, S., C. Blake, and S.L. West. *Information Extraction from Medical Notes (P187)*. in *12th World Congress on Health (Medical) Informatics*. 2007. Brisbane, Australia: MedInfo.
25. West, S.L., S.B. L., and P. C., *Validity of Pharmacoepidemiology Drug and Diagnosis Data in Pharmacoepidemiology* S.B. L., Editor. 2005, John Wiley and Sons, Ltd: Sussex. p. 709-765.
26. Shih, A., et al., *Organizing the U.S. Health care Delivery System for High Performance*. 2008, The Commonwealth Fund. [accessed 2008 September 5]; Available from: [http://www.commonwealthfund.org/publications/publications\\_show.htm?doc\\_id=698139](http://www.commonwealthfund.org/publications/publications_show.htm?doc_id=698139)

27. Kokko, S., *Towards fragmentation of general practice and primary healthcare in Finland?* Scand J Prim Health Care, 2007. **25**(3): p. 131-2.
28. Clarfield, A.M., H. Bergman, and R. Kane, *Fragmentation of care for frail older people--an international problem. Experience from three countries: Israel, Canada, and the United States.* J Am Geriatr Soc, 2001. **49**(12): p. 1714-21.
29. Detmer, D. and S. E. *Learning from Abroad: Lessons and Questions on Personal Health Records for National Policy.* 2006 [cited 2007 December 20]; Available from: [http://www.esi-bethesda.com/ncrrworkshops/clinicalResearch/pdf/2006\\_10\\_phr\\_abroad\\_DED\\_AA\\_RP.pdf](http://www.esi-bethesda.com/ncrrworkshops/clinicalResearch/pdf/2006_10_phr_abroad_DED_AA_RP.pdf).
30. Working Group on Accurately Linking Information for Health Care Quality and Safety (2005) *Linking Healthcare Information: Proposed Methods for Improving Care and Protecting Privacy. Volume,*
31. American Association for Health Freedom. *Position Paper: Medical Privacy.* 2006 [accessed 2007 December 20]; Available from: <http://www.apma.net/aahf/showarticlenew.asp?articleid=49>.

## Figure Legends

**Figure 1** WebCIS files used for the diabetes case study with medical record number as the primary linkage variable

**Figure 2** Examples of quality indicators. ACE = angiotensin-converting enzyme

**Figure 3** De-identification procedure using a hash table of associative arrays containing the first name (FName), last name (LName), medical record number (MRNo), and SSN represented by a pointer (ptr).

**Figure 4** Patient identification process. DM = diabetes mellitus type 2; HbA1c = hemoglobin A1c level.

**Table 1** Clinical and administrative areas contained in the current WebCIS version

• Allergy lists	• Pathology reports
• Cardiology reports	• Patient appointment scheduling
• Clinic visit data	• Patient demographic data
• Electrocardiogram results	• Patient problem file
• Gastrointestinal procedures reports	• Peripheral vascular laboratory tests
• Hospital census information	• Pulmonary reports
• Insurance provider	• Radiology reports
• Laboratory results	• Referring physician lists
• Medical imaging	• Respiratory therapy reports
• Medications prescribed	• Transcribed notes

**Table 2** Exclusion and inclusion criteria for identifying patients with newly diagnosed diabetes mellitus type 2

<b>Set 1</b>	<b>Set 2</b>
<p><u>Exclusion criteria:</u></p> <ul style="list-style-type: none"> <li>• ICD-9-CM diagnosis code of 250.xx in the patient problem file before January 1, 2001, and/or</li> <li>• Medication prescribed for diabetes mellitus before January 1, 2001</li> </ul>	<p><u>Exclusion criteria:</u></p> <ul style="list-style-type: none"> <li>• Mention of diabetes type 1 or type 2 before January 1, 2001</li> <li>• Steroid-induced diabetes</li> <li>• Control of diabetes through diet alone</li> </ul>
<p><u>Inclusion criteria:</u></p> <ul style="list-style-type: none"> <li>• ICD-9-CM diagnosis code of 250.xx in the patient problem file after January 1, 2001</li> <li>• <math>\geq 2</math> HbA1c laboratory tests after January 1, 2001</li> <li>• <math>\geq 2</math> outpatient visits in the 2-year period before the first HbA1c laboratory test at UNC</li> <li>• Elevated HbA1c level after January 1, 2001</li> <li>• Antidiabetic medication prescribed after January 1, 2001</li> </ul>	<p><u>Inclusion criteria:</u></p> <ul style="list-style-type: none"> <li>• No mention of diabetes before January 1, 2001</li> <li>• Statement that patient did not have diabetes</li> <li>• Hyperinsulinemia with or without metformin treatment</li> <li>• Borderline or questionable diabetes (i.e., firm diagnosis not made)</li> <li>• Polycystic ovary syndrome</li> <li>• Gestational diabetes</li> </ul>