North American Association of Central Cancer Registries, Inc. (NAACCR)

Discharge and Claims Data Best Practices Guide

October 2015

Acknowledgements

This activity is supported in part by Cooperative Agreement Number 1U58DP004917-02 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

Suggested Citation: NAACCR Discharge Data Work Group (eds). Discharge Data Best Practices Guide. Springfield, IL: North American Association of Central Cancer Registries, October, 2015.

Table of Contents

Discharge and Claims Data Best Practices Guide Editors

Dan Curran, MS, CTR
C/NET Solutions
Executive Editor

Sandy Jones
Sections 1 and 5: *Introduction and Gaining Access to Discharge and All Payer Claims Data*
Cancer Surveillance Branch
Division of Cancer Prevention and Control
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention

Patricia Andrews, MPH, CTR
Section 2: *Uses of Discharge and Claims Data in the Central Cancer Registry*
Louisiana Tumor Registry
Louisiana Health Sciences Center, School of Public Health Epidemiology Program

Saba Yemane, BA, BS
Section 3: *Linkage Software and State Experiences*
Missouri Cancer Registry and Research Center
University of Missouri School of Medicine
Department of Health Management and Informatics

Mark Cruz, CTR
Section 4: *Overcoming Obstacles*
Mercy Cancer Institute, Sacramento

Discharge Data Best Practices Guide Reviewers

Patricia Andrews, MPH, CTR
Louisiana Tumor Registry
Phone: 504-568-5795
E-mail: pandre@lsuhsc.edu

Annette A. Hurlbut, RHIT, CTR
Elekta Impac Software
Phone: 315-593-7130
E-mail: annette.hurlbut@elekta.com

J. Jackson-Thompson, MSPH, PhD
Missouri Cancer Registry and Research Center
Phone: 573-882-7775
E-mail: jacksonthompsonj@health.missouri.edu

Cyllene R. Morris, DVM, PhD
California Cancer Registry
Phone: 916-731-2503
E-mail: cmorris@ccr.ca.gov

Brent J. Mumphrey
Louisiana Tumor Registry
Phone: 504-568-5767
E-mail: bmumph@lsuhsc.edu

NAACCR Discharge Data Work Group Roster

Dan Curran, MS, CTR, Chair
Public Health Institute - C/NET Solutions
Phone: 916-779-0362
E-mail: danc@askcnet.org

Patricia Andrews, MPH, CTR
Louisiana Tumor Registry
Phone: 504-568-5795
E-mail: pandre@lsuhsc.edu

Mark Cruz, CTR
Mercy Cancer Institute
Phone: 916-537-5069
E-mail: Mark.cruz@dignityhealth.org

Lori A. Havener, CTR
NAACCR
Phone: 217-698-0800 ext. 3
E-mail: lhavener@naaccr.org

Qiming He, PhD
ICF Macro
Phone: 301-272-8527
E-mail: qiming.he@macrointernational.com

Annette A. Hurlbut, RHIT, CTR
Elekta Impac Software
Phone: 315-593-7130
E-mail: annette.hurlbut@elekta.com

Sandy Jones
Cancer Surveillance Branch
Division of Cancer Prevention and Control
Centers for Disease Control and Prevention
Phone: 770-488-5689
E-mail: SFT1@cdc.gov

Brent J. Mumphrey
Louisiana Tumor Registry
Phone: 504-568-5767
E-mail: bmumph@lsuhsc.edu

Sumana Nagaraj
North Carolina Central Cancer Registry
Phone: 919-715-9728
E-mail: Sumana.Nagaraj@dhhs.nc.gov

Serban Negoita, MD, DrPH, CPH, CTR
Maryland Cancer Registry
Phone: 240-314-2309
E-mail: SerbanNegoita@westat.com

David Stinchcomb
Westat
Phone: 301-610-5571
E-mail: DavidStinchcomb@westat.com

Saba Yemane
Missouri Cancer Registry and Research Center
University of Missouri School of Medicine
Department of Health Management and Informatics
Phone: 573-884-6130
E-mail: yemanes@health.missouri.edu

Section 1: Introduction

By Sandy Jones

Hospital discharge data (HDD), all-payer claims data (APCD), and cancer registry data have typically been used as separate resources for understanding cancer prevalence, incidence, resource use, quality of care, and cost. Opportunities exist to link these sources to further address important questions for cancer researchers, epidemiologists, health services researchers, policy makers, and program managers.

The purpose of the discharge data project was twofold. First, the Centers for Disease Control and Prevention (CDC) supported the National Association of Health Data Organizations (NAHDO) to conduct a gap analysis to identify data elements that could be standardized, added, or improved across these data sources. Second, an examination of the types of questions that could be answered using combined databases was conducted. State agency and federal staff provided NAHDO with input on the gaps, priorities, and limitations of the data sources.

The North American Association of Central Cancer Registries, Inc. (NAACCR) developed this *Discharge and Claims Data Best Practices Guide* to provide information to central cancer registries (CCRs) about linkage methodologies and uses of HDD and APCD by cancer registries.

Background

The HDD and APCD are rich sources of data that can provide cancer registries with information to fulfill their reporting requirements. However, reporting requirements for HDD, APCD, and cancer registries have never been harmonized at the national level. Hospital discharge data (an abstract of information from the UB (Uniform Billing)-04 claims information)[1] have been collected and widely used for more than 30 years. For example, The Healthcare Cost and Utilization Project[2] (HCUP) data, from the Agency for Healthcare Research and Quality (AHRQ), is a family of healthcare databases and related tools for research and decision making. The databases contain a core set of clinical and non-clinical information derived from the data collection efforts of organizations in participating states that maintain statewide data systems. Even though these data have been used extensively, weaknesses in the data remain and present challenges to CCRs for linking and using the resulting linked data.

The state HDD reporting systems include data from inpatient discharge abstracts, hospital-affiliated emergency department abstracts, and ambulatory care encounters in hospital-affiliated (and sometimes freestanding) ambulatory surgery sites. Currently, all states except Alabama, Delaware, and Idaho but including the District of Columbia require standard reporting of HDD to their public health department or designee organization. The state APCD reporting system includes data from all payers for inpatient, ambulatory surgery, emergency room, outpatient, clinic, pharmacy, and physician office services. At least 12 states have implemented statewide APCD reporting systems and more states are moving to implement systems of this type.[3] Data from these two systems are reported in a standard format on a routine basis to the state health department. The state-based American Health Information Management Association (AHIMA) offices oversee medical record coding standards and are responsible for implementing the hospital discharge and claims data

reporting systems in their respective states. NAHDO provides technical support and overall direction to states on the implementation of these systems.

Why ANSI X12?

State agencies that maintain central cancer registries have relied primarily on clinical information from hospital medical records to obtain data for mandated cancer surveillance reports. Changes such as the migration of cancer diagnosis and treatment to outpatient facilities, the need to be cost effective by obtaining data from new sources, and the standardization of electronic reporting have driven an examination of alternative data sources.

Obtaining cancer diagnoses directly from electronic pathology reports is one example of enhancing efficient data ascertainment. Another potential data source for cancer reporting is insurance claims data that are submitted electronically in a standardized format. The purpose of this guide is to provide guidance for the use of hospital discharge and insurance claims data for cancer surveillance reports.

The use of the American National Standards Institute (ANSI) X12 standard for health insurance data was mandated by the Health Insurance Portability and Accountability Act (HIPAA) enacted in 1996. HIPAA required the use of standard transaction formats and code sets for the electronic transmission of health information under rules promulgated by the federal Department of Health and Human Services. HIPAA designated the Accredited Standards Committee (ASC) X12, a standard-developing organization accredited by ANSI, as the standard setters for health insurance data. Specifically, the X12N-Insurance Subcommittee creates and maintains standard data elements for healthcare transactions such as claims and encounters, eligibility inquiries and responses, claim status inquiries and responses, referrals and prior authorizations, and healthcare payment and remittance advice. The ASC X12N-Insurance body meets three times per year to develop and maintain standards.

The transaction names and identifiers that are important for cancer surveillance use include:

- 837 Claims
- 837 Health Care Services Data Reporting Guide

The HIPAA Electronic Transaction Standards and implementation dates are as follows:

- Version 4010 – October 1997 – Previous standard
- Version 5010 – October 2003 – Adopted in 2013 and updated for ICD-10
- Version 6020 – Reviewed in public comment period – adoption timeline to be determined for possible future implementation

Standards designated by X12N-Insurance include:

- International Classification of Diseases (ICD) diagnosis and procedure codes
- National Drug Codes
- Health Care Financing Administration's (HCFA's) Healthcare Common Procedure Coding
- System (HCPCS) and Current Procedural Terminology, Fourth Edition (CPT-4) for Physician and Other Services
- Preferred Language Spoken
- Source of Payment Typology

- Present on Admission
- Office of Management and Budget (OMB) Race and Ethnicity Codes

Sources of claims data include:

- Medicare
- Medicaid
- All Payer Claims Databases
- AHRQ's HCUP

Claims data have been widely used in health departments, but not typically for cancer surveillance. An assessment of claims data has been completed using cancer registry data as the gold standard and focusing on certain cancers that are underreported in the hospital medical record. Several state cancer registry programs have implemented processes to link their state population cancer registry data with Medicare data.[4] The Surveillance, Epidemiology and End Results (SEER) Program dataset linking a sample of the SEER cancer cases with Medicare data is available to researchers.

Overview of the X12 Standard

The ASC X12 standards developed by the ASC X12 group are widely used across a wide range of industries including healthcare, insurance, transportation, and finance to streamline business transactions and use a common, uniform business language.[5]

HIPAA mandates that Uniform Billing (UB) information systems must use the ANSI ASC X12 837 claim/encounter standard. Most states are utilizing the UB data as a source to report their discharge and claims information, eliminating the need for duplicate data collection and implementing uniform standards to enable interoperable data exchange. Medicaid Managed Care is another user of the UB data specifications. The latest version of the UB is UB-04, which replaced the UB-92.[6]

Discharge and Claims Data Content Information

In some state discharge databases, Social Security Numbers (SSNs) are available to assist with linkage. However, many states have statutory restrictions preventing use of the SSN as a patient identifier; under these restrictions, hospital discharge systems use a state-assigned patient identifier. This state-assigned identifier is generally useful only for the state in which it is assigned and is not as effective as an SSN, which is considered a "strong identifier." Yet the lack of a strong identifier does not preclude linkage, given other variables in the data that can be used in a probabilistic matching program. Linking individual cases using probabilistic matching, however, does result in some failed matches due to either coding errors or case assignment errors in the data.

Analysis of combined data linked by strong identifiers allows researchers to answer more questions than linkages made with weaker identifiers allow. For example, when SSN or other strong patient identifiers are included, questions related to readmissions can be readily answered because tracking of hospital stays across hospitals can be completed. Unfortunately, only readmissions to the same hospital can be tracked when using weaker identifiers, although those states with strong privacy rules may allow the discharge system to create a readmission flag indicating whether the specific discharge was a readmission for a prior hospitalization.

Those looking for extensive clinical information will find only a limited amount in discharge data sources. Primary and secondary diagnoses are identified using ICD codes. Included with these clinical diagnoses are "present on admission" codes, identifying whether the condition was present at the point of admission to the hospital (comorbidities in cancer registry abstracts). Comparison of these codes with discharge codes allows for the identification of hospital-acquired conditions such as infections. Procedure coding is more robust given that hospitals are paid on the basis of the procedures performed. Procedures are identified using either ICD-9 procedure codes or CPT codes. Other related data elements, such as discharge status codes, allow for determination of whether the patient died in the hospital, was released to home, or transferred to another hospital or facility.

Correctly used procedure codes are strong data elements because these codes impact payment. For some codes, discharge data may be equivalent, or even superior, to clinical record information. However, diagnostic codes that indicate high severity (thereby increasing payment) should be viewed with caution if not supported by clinical information due to the possibility of "upcoding" - assigning an inaccurate billing code to (a medical procedure or treatment) to increase reimbursement - to enhance revenue.[7]

Some key elements commonly found in cancer registries will enhance the clinical values that are present in discharge data. For example, staging and site of tumors, found in registry data, could enhance the information found in the discharge data. This could improve the risk adjustment methods for cancer resource use analyses. It also allows for more detailed utilization and disparities analyses. Discharge databases provide information on length of stay and accompanying hospital charges. The APCD is a more robust system of data that includes inpatient and outpatient services. APCDs often include actual payment information as well as charges, allowing the determination of the real cost for the payer. Both discharge data and all-payer data include detailed payer type information, but generally only hospital discharge data includes cases where payment was made by the patient (identified as "self pay").

Data users need to carefully think through the best method for getting at the most abundant and meaningful data found in each of the databases.

NAACCR and Discharge Data (UB-04) Crosswalk: A payer typology crosswalk between the cancer registry data dictionary and the discharge data dictionary was developed by NAHDO, CDC, and NAACCR members. The crosswalk will be published to http://www.naaccr.org/StandardsandRegistryOperations/InteropInfo.aspx when approved.

Several of the recommendations have been addressed through a variety of ways, such as the publishing of a joint statement between NAHDO and NAACCR regarding the importance of a national identifier and standardized personal identification information that will improve linkages between multiple data sources, posted on both the NAHDO and NAACCR websites. Other recommendations have been addressed through educational webinars provided to both the cancer registry and the discharge data communities. The joint statement and recorded webinars are posted on the NAACCR website at: http://www.naaccr.org/StandardsandRegistryOperations/InteropInfo.aspx.

References

1. CMS Manual System Pub 100-04 Medicare Claims Processing https://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/Downloads/R1104CP.pdf

2. Hospital Cost and Utilization Project (HCUP): http://www.ahrq.gov/research/findings/factsheets/tools/hcupdata/index.html.

3. The All-Payer Claims Database Council: http://www.apcdcouncil.org/.

4. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. Agreement of Medicare claims and tumor registry data for assessment of cancer-related treatment. *Med Care*. 2000 Apr;38(4):411-21.

5. The Accredited Standards Committee X12: http://www.x12.org/about/index.cfm.

6. Public Health Data Standards Consortium: http://www.phdsc.org/standards/x12/aseds.asp.

7. Li, B. Cracking the Codes: Do Electronic Medical Records Facilitate Hospital Revenue Enhancement? http://www.kellogg.northwestern.edu/faculty/b-li/JMP.pdf Kellogg School of Management, Northwestern University.

Section 2: Uses of Discharge and Claims Data in the Central Cancer Registry

By Patricia Andrews, MPH, CTR

Hospital discharge data (HDD) can supplement traditional central cancer registry (CCR) efforts to identify cases; update follow-up records; and enhance treatment, comorbidity, and insurance information. In Canada, the comparable source of information is known as the Discharge Abstract Database, but this summary will focus on the HDD because more published information is available for HDD.

HDD is now available in most states—some states limit the data to hospital inpatients while others include ambulatory/outpatient facilities. The availability of patient identifiers in HDD varies from state to state; where they are missing, linkages are not possible. Some of the uses of HDD data from linkages are described below.

Case Identification

Supplementing traditional cancer registry case-finding techniques with records from hospital discharge data will increase the number of cases in a registry. Because the Florida Cancer Data System (FCDS) has found HDD linkages useful, it incorporates linkages as a routine year-end procedure and considers them an "excellent case-finding tool."[1]

The Louisiana Tumor Registry recently linked its database with state HDD to evaluate its identification of benign and borderline brain cancers from 2004 through 2011. After a time-consuming review of the cases listed in the HDD but not in the registry database, 46.5 percent were determined to be reportable but missed. This valuable information was used to evaluate and revise case-finding strategies for non-malignant brain cancers.

Linkages of CCR and HDD databases do have shortcomings, however. For example, HDD records will identify many cases absent from the registry, but the experience of the New York State Cancer Registry has shown that most HDD "cases" are already in the registry or are not reportable.[2] The New York Cancer Registry has found HDD to be a valuable source of benign brain cases, although "most" HDD cases that are followed back are not reportable. Two studies in Italy found that HDD, when used alone, missed more than 10 percent of the cases identified through routine registry operations.[3,4]

Screening the Present on Admission (comorbidities) section of HDD will identify patients whose cancers were missed by routine hospital casefinding but who were admitted to a hospital because of other conditions with cancer mentioned as a pre-existing condition. Similarly, complications from cancer-related procedures may identify reportable cases that had been treated only in non-hospital settings. On the other hand, following back on these leads may waste considerable time to eliminate prevalent cases that are already in the registry.

A more fruitful source of identifying previously missed cases is APCD, which is a compilation of claims from hospitals, physicians, and clinics to insurance companies. While hospital-based

screening will miss the cancer diagnoses of patients treated only in clinics and physician offices, such as elderly prostate cancer patients, physician billing records will definitely include them and will provide detailed treatment information. Self-pay services as well as those for uninsured patients are excluded from the APCD. More than 20 states compile or are implementing APCD (http://www.apcdcouncil.org/state/map).

Through a linkage of registry data with five hematology-oncology practices in North Carolina and Virginia, Lynne Penberthy et al. determined that only 56 percent of lymphohematopoietic cases treated in the clinics had been identified by the registries.[5] (Additional details on Dr. Penberthy's use of linkages with APCD can be found in the next section.) Drawbacks of using APCD include the failure of most states to compile the statistics and the potential time wasted spent ruling out ineligible or prevalent cases.

Treatment Information

Many of the procedures listed in HDD and APCD will be for treatment rather than diagnosis, and both databases will provide details of treatment after the 6 months traditionally captured in routine abstracts. The accuracy of APCD billing data varies by type of cancer but usually exceeds 95 percent,[5] although this information source does not capture most orally administered agents.

During a 2012 NAACCR webinar, the FCDS recommended HDD as a source of surgical information.[1] A Nebraska linkage of registry and HDD data increased the number of cases with recorded radiation treatment by at least 12 percent for colorectal cancer cases and 15 percent for breast cancer; the additional data applied predominantly to rural residents.[6] During the same NAACCR webinar, Penberthy, a leader in the use of APCD, pointed out that it contains the detailed information that is essential for comparative effectiveness research.[5] Her investigation in California increased treatment information by 29 percent for radiation, 32 percent for chemotherapy, 12 percent for hormonal treatment, and 44 percent for biological response modification.

Taking advantage of the information from linkages with HDD and APCD, however, will require considerable staff time and judgment before incorporating treatment data into abstracts. For example, if a person has more than one cancer, for which case is a treatment administered?

Comorbidities

HDD may document a number of different comorbid conditions, called Present on Admission conditions, as well as complications. States vary in the number they allow. These coded diagnoses can be added directly to the relevant section of an abstract, but the registry will need to avoid duplicating conditions already listed in the abstract.

Comorbidity data are a valuable resource for researchers correlating cancer treatment decisions with pre-existing conditions or investigating diseases that might be a risk factor for cancer. A California study, however, noted that HDD are more likely to include acute than chronic conditions and that cancer may be overlooked as Present on Admission if the person has been admitted for a totally different reason.[9]

Case Follow-Up

HDD can provide follow-up dates for cancer patients who go to a hospital for any reason. The FCDS has found HDD an excellent source of follow-up dates of last contact[1] while the Louisiana Tumor Registry found new dates of last contact for cases aged less than 20, a traditionally difficult age group to follow

Death Certificate Follow-Back

Some states are finding HDD to be a useful resource for investigating deaths attributed to cancer in death certificate only (DCO) cases. The New York Cancer Registry reports that HDD linkages resolve almost all its potential DCO cases. [2] The Missouri Cancer Registry and Research Center (MCR-ARC) submits the identifiers of unresolved potential DCO cases to the state health department after exhausting all feasible traditional methods of obtaining information. Linkage of MCR-ARC's list to the HDD results in matches for the majority of potential DCO reports. After a Certified Tumor Registrar (CTR) review and a comparison of the HDD diagnosis codes to the type of cancer on the death certificate, many of the potential DCO abstracts may be resolved and converted to non-DCO cases using the earliest HDD date for that disease as the diagnosis date. Another outcome is that cases may be deemed non-reportable. The HDD linkage resolved approximately 30-45 percent of Missouri's potential DCO abstracts for reporting years 2010 and 2011. For reporting year 2010 alone, this process reduced the number of overall DCO cases by 37 percent relative to what it would have been if not for the HDD linkage (this linkage is also described in Section 3).

Race

The New York State Cancer Registry, which has several years' experience with HDD linkages, reports that this procedure provides information for about 40 percent of the cases that lack race.[2] Missouri has also found the linkage informative for race, particularly Asians and Pacific Islanders, and for Hispanic ethnicity data.[7]

Primary Payer

Insurance is important to hospitals, so virtually all HDD reports include information on this topic. The New Jersey Cancer Registry reported a 50 percent decline in cases coded to "Unknown" for primary payer after linkage with HDD.

Some caveats apply to the use of HDD insurance information. HDD allows the inclusion of multiple payers while most registries limit the number of payers or combinations of payers. Thus, registries must develop policies about incorporating HDD insurance data into existing abstracts. Since HDD covers all hospital visits during and after a patient's diagnosis, insurance information may include coverage that post-dates diagnosis. Because cancer registries are supposed to record primary payer only *at diagnosis*, they will need to restrict HDD insurance information to entries that match or closely follow the diagnosis date. If HDD includes multiple insurers at diagnosis, determining which is primary is very difficult.

Because types of insurance coverage may be recorded in several different coding systems in hospital records and HDD, the NAACCR Semantic Interoperability Workgroup is preparing a crosswalk for converting these into NAACCR codes. A re-abstraction study for insurance showed that approximately 85 percent of HDD primary payer data agreed with that in the cancer registry data.[8]

Stage and Histology

These data items are missing from HDD and APCD.

Research

Data from HDD have been linked with those from trauma and cancer registries, environmental tracking data, the American Hospital Association, and other databases. The New York State Cancer Registry has used the linked database for studies of hospital length of stay, outpatient cancer surgery, and the effect of comorbidities on survival.[2]

A caveat is that definitions for the HDD variables can vary from state to state and may differ from their apparent NAACCR counterparts. Researchers must determine the comparability of data from various sources. At the same time, both HDD systems and NAACCR should work to improve the interoperability of their data items so that these rich sources of data can be combined accurately.

Summary

HDD compiles large amounts of information that can be used to improve the quality of cancer registry data. By the nature of its sources and its development, however, HDD need to be understood well in order to add them to cancer registries efficiently and accurately.

References

1. MacKinnon JA. Utilization of statewide discharge data in Florida. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

2. Boscoe F, NAACCR discharge data webinar. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

3. Ferretti S, Guzzinati S, Zambon P, Manneschi G, Crocetti E, Falcini F, Giorgetti S, Cirilli C, Pirani M, Mangone L, Di Felice E, Del Lisi V, Sgargi P, Buzzoni C, Russo A, Paci E.[Cancer incidence estimation by hospital discharge flow as compared with cancer registries data]. *Epidemiol Prev* 33(4-5):147-53. 2009.

4. Yuen EJ, Louis DZ, Rabinowitz C, Maio V, Cisbani L, DePalma R, Grilli R. Using hospital discharge abstract data to identify incident breast cancer cases and assess quality of care. *BMC Health Services Research,* 10 (supple 2):A1. http://www.biomedcentral.com/1472-6963/10/S2/A1.

5. Penberthy L. Using claims data for automated case finding and treatment reporting from multiple community specialty practices. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

6. Lin G. Enhance cancer care surveillance by linking cancer registry and hospital discharge data. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

7. Jackson-Thompson J, Schmaltz C. Missouri Cancer Registry and Research Center (MCR-ARC): improving data quality and completeness through Patient Abstract System (PAS) (=MO hospital discharge data) linkage. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

8. Verrill C. "Assessing the Reliability and Validity of Primary Payer Information in Central Cancer Registry Data," NAACCR conference 2010.

9. Morris, CR. Use of hospital discharge files to enhance registry data: California experience. Presentation at NAACR webinar, Discharge Data Implementation: Challenges and Rewards, Oct 10, 2012.

Section 3: Linkage Software and State Experiences

By Kathleen K. Thoburn, CTR and Saba Yemane

Linkage Software Used by the Majority of Cancer Registries

Efficient, accurate record linkage is a vital activity for all central cancer registries (CCRs). Registries conduct record linkage on a daily basis for casefinding, linking new reports of cancer to the CCR, detecting duplicate cases, conducting follow-up, and using external data files for special studies. Failure in patient linkage processes results in missed cases and/or duplicate registrations, missed information (e.g., vital status) from other data sources, and generation and reporting of inaccurate cancer counts and rates.[1] As a result, the selection of linkage software is an important decision for any cancer registry. Two software programs commonly used by CCRs are Link Plus and AutoMatch.

Note: Although this document is titled *Discharge and Claims Data Best Practices Guide,* due to the importance of record linkage to CCRs and the fact that linkage with hospital discharge data (HDD) is a new or relatively new activity for many registries, the authors of this section feel justified in including descriptions of other uses for linking software. Therefore, several states' experiences in linking with other databases are included in Appendix A.

Link Plus

Link Plus is a probabilistic record linkage program included in the Registry Plus™ suite of software developed by CDC's Division of Cancer Prevention and Control in support of CDC's National Program of Cancer Registries (NPCR). Link Plus is a stand-alone, probabilistic record linkage program that combines user-friendly interfaces with complex statistics. The easy-to-use program is available free of charge from the CDC NPCR website: http://www.cdc.gov/cancer/npcr/. Link Plus accepts fixed-width and delimited files; it can be used to detect duplicates within a database or to link cancer registry files with external data files. Although originally designed for use by cancer registries, the program can be used with any type of data files in fixed-width or delimited format.

Link Plus accommodates linkages between significantly large data files, allows one-to-one matching, many-to-many matching, optional generation of a non-match report, and multiple export options for linkage results, including export of merged data in NAACCR file format. Link Plus also utilizes OR blocking methods, and is programmed to eliminate the need for multiple passes when conducting a linkage. In addition, the program includes pre-configured linkage with Breast and Cervical Cancer Early Detection Program data.[2] The latest version of Link Plus (version 3.0) is still in beta mode, but can be obtained upon request by contacting cancerinfo@cdc.gov.

AutoMatch

The most widely known commercial probabilistic record linkage program is AutoMatch, developed by MatchWare Technologies Inc., Silver Spring, MD. The last version of AutoMatch (4.2) was released in 1992. AutoMatch utilizes OR blocking methods and requires multiple passes when conducting linkages. AutoMatch was first developed in a DOS version and later was converted into a Windows Graphical User Interface (GUI) called Essential. Essential is expensive, requires annual license renewals, and is far beyond the scope of many research groups. As a result, AutoMatch is no

longer widely used in research contexts but is often used for validation of other linkage programs. Only a few cancer registries use the old DOS version of AutoMatch with special permission of IBM.[3]

References

1. Thoburn KK, Gu D, Rawson T, and Rogers JD. Link Plus Version 2: An Essential Central Cancer Registry (CCR) Linkage Tool. *North American Association of Central Cancer Registries*, June, 2008.

2. Thoburn KK, Gu D, Rawson T, and Rogers JD. Link Plus Version 3: Overview of Enhancements. *North American Association of Central Cancer Registries*, June, 2009.

3. Schnell, Rainer, Record-Linkage from a Technical Point of View (August 2009). RatSWD_WP_ 124. Available at SSRN: http://ssrn.com/abstract=1462075 or http://dx.doi.org/10.2139/ssrn.1462075.

Section 4: Overcoming Obstacles

By Mark Cruz, CTR

This brief overview identifies some of the obstacles associated with hospital discharge data (HDD) and highlights some solutions to overcome them.

Patient identifiers such as social security number (SSN) and date of birth (DOB) are among the most important items required to ensure the correct data is matched to the correct patient in this new era of electronic interchange. With an increasing number of facilities masking or partially masking SSN and DOB (e.g., including only the last four digits of the SSN or the year of birth) in order to protect patients' private health information and minimize the risk of potential data breaches, this decision presents a formidable obstacle to ensuring data accuracy. This has required central cancer registries (CCRs) to utilize different options to obtain accurate patient identifiers. For example, registries use internet resources such as Masterfiles[1] to properly identify patient demographics. Another option is to use a vendor service such as Ingenix (now known as OptumInsight)[2] to mine the internet for patient information. Vendors can customize searches to help minimize the costs and provide the data necessary to augment a registry database. The HIPAA exemption from the patient consent requirement for cancer reporting is a major asset when trying to obtain patient identifiers and engaging with vendors for their services.

Physician offices are a cornerstone to obtaining patient information but can be challenging; office staff are often not educated on cancer reporting requirements or cannot comply with the timeline for providing data to the registry. Staff turnover is also a concern; CCRs may invest a considerable amount of time in educating specific staff members but when they leave, the process has to start all over again. One of the process improvements undertaken by the Cancer Registry of Greater California (CRGC) has been to engage with the practice administrators and have them be the point of contact for physician reporting requests. Media kits specifically created to inform physicians and their staff have also been employed with success. These efforts have increased the number of self-reporting physicians within CRGC.

Veterans Administration (VA) facilities can present a bottleneck for obtaining or accessing data. Because these facilities operate under federal authority, state laws do not apply. CCRs can, however, negotiate data use agreements (DUAs) with VA facilities in their state that allows a VA facility to send its cancer cases to the state registry. Often, VA cases will link to an admission to, or treatment by, a non-federal facility, allowing for a more complete picture of the patient's cancer case and treatment.

The availability of CTRs in VA facilities is another obstacle some CCRs face. The VA employee assigned to the registry may not be a CTR and may be unaware of reporting requirements or available resources. One way to overcome this is to develop a mentor program so that a designated state registry employee can help the employee at the VA facility better understand the role of the CTR. Another way the CCR can facilitate reporting by VA facilities is to invite VA registrars and their supervisors to participate in workshops and registry trainings.

Variation in state legislation represents another challenge. While federal funders and standard setters require reporting of specific data elements using standardized codes in standardized layouts, the

method of reporting cancer cases and compliance requirements are determined at the state level through laws, statutes, and regulations. Some states have been very proactive with cancer reporting legislation and have enacted laws that give the central registry significant muscle to obtain the necessary cancer data. Other states' cancer legislation is less robust. Where penalties are infractions that have few financial or other consequences and CCRs have limited staff and resources, non-hospital reporting is likely to be incomplete. A uniform federal cancer reporting law could eliminate confusion on access to cancer information; reduce the time spent on educating reporting sources; and increase the completeness, timeliness and quality of the data. Any potential uniform legislation should address emerging electronic data transfer capabilities and Meaningful Use (MU) implementations.

What impact do claims data limitations have on the availability of discharge data? We know that timely submission of claims is critical to the medical provider's revenue cycle. Successful monitoring of the claim lifecycle is an important element to having quality data available for multiple uses. If a claim is denied, the provider's staff must go through the appeals process, keeping the data from being available in a timely fashion for other uses. The completeness of the data fields on the claim are also an important element to having detailed data available for cancer registries. The lack of complete and accurate recording of the data can result in insufficient cancer data being available in the insurance companies' database for successful data linkages. The effectiveness of the clearinghouse or insurance company to successfully share their data in a nationally accepted data transfer format is a major obstacle in trying to access and properly use all payer claims data (APCD) as is incomplete adoption of APCD by states. The Centers for Medicare and Medicaid Services (CMS) recommends a layout but each software vendor can modify it to suit their specifications and programming for transmit to a clearing house or insurance provider. Most want to be compliant with CMS' suggestions, so a single format for usage would be a significant incentive, but the cost of implementation would be another obstacle. With MU Stage 2 incentives for eligible providers and eligible hospitals, an increase in claims data timeliness can be expected for some providers who may hold their billing for various reasons (e.g., if it is close to the end of the month and they have hit the capitation max or are waiting to bundle a claim). This may change that practice and result in faster submissions. Also, the information coded on the claim form will benefit not only central cancer registries, but all entities involved in data exchange and outcomes. Additional data may be able to be collected by research and surveillance organizations.

The availability of software and the costs associated with implementation are other obstacles to consider. Oftentimes, the standard setters (SEER, NPCR, American Joint Committee on Cancer, and Commission on Cancer) may not release critical information that is required to make programming or specification updates to software in a timely manner. These delays then force software vendors to delay their release maps to cancer reporting facilities. In turn, the facility must delay routine reporting, resulting in unfavorable timeliness, completeness, and/or quality assurance edits. The facility notifies the regional/central registry when the vendor will have a delay in a software release in an effort to mitigate the impact on the facility (setting an exclusion flag on the delayed cases so the facility is not penalized). The regional/central registry also monitors these situations, as delays in software releases can cause a backlog in submissions to the regional/central registry and can impact the regional/central data transmissions to their stakeholders. The registry-related costs of the software, IT support, and maintenance to healthcare facilities also are obstacles to consider. More and more facilities opt out of certain national programs, affiliations, and collaborative efforts as federal, state, and local funding is reduced; the elimination of facility-based cancer registries or the

hiring of non-CTRs may be proposed by facility administration as a cost-savings option. Comprehensive communication and adherence to an agreed-upon timeline will help avoid these obstacles and will result in a better collaborative relationship among standard setters, vendors, facilities, and regional/central registries.

References

1. Master Files. http://www.masterfiles.com/ accessed August 14, 2015.

2. Ingenix name retired as United re-brands subsidiaries. http://www.amednews.com/article/20110426/business/304269998/8/ accessed August 14, 2015.

Section 5: Gaining Access to Discharge and All Payer Claims Data

By Sandy Jones

The release of identifiable and de-identifiable discharge data is governed by state and federal laws but the stewards of discharge data may be employed by public or private entities such as hospital associations. One of the following models currently exists in states:

1. State agency with legislative mandate for collecting data;
2. Delegated authority, such as a hospital association or private entity, collecting data under a state mandate; or
3. Private agency, usually a hospital association, collecting the data voluntarily from its members or community hospitals

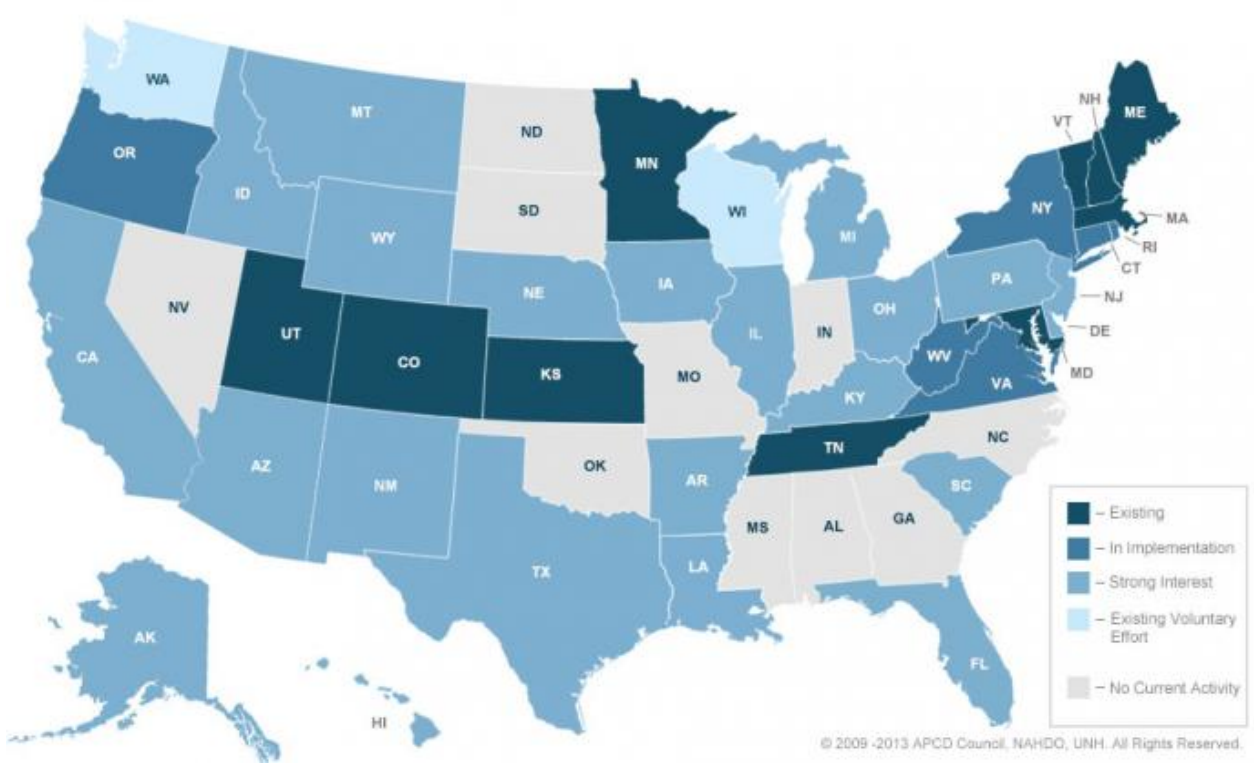
Data disclosure policies vary across states that have a private agency collecting the data either voluntarily or under a state mandate. Establishing a personal relationship with the data steward is critical to successfully gain access to the data. The data stewards must clearly understand a project and feel that they are included in discussions and decisions about how the data will be utilized. Some states charge a fee for the data to offset their data collection costs, or the data steward may accept analytic support or tools in lieu of money.

If a cancer registry program is having difficulty accessing state discharge data, then the National Association of Health Data Organization (NAHDO) may be able to facilitate a partnership with the state steward of the state discharge data set. Visit the NAHDO website for more information on how NAHDO works with state public health agencies to move toward standardized discharge and claims systems that will be interoperable across the country.[1]

The NAACCR Discharge Data Workgroup compiled information from state cancer registries through a NAACCR survey completed in 2011. NAACCR received a total of 38 responses, which included 31 states, Greater Bay Area Cancer Registry in California, Puerto Rico, Guam, Ontario, Quebec, Saskatchewan, and Alberta. Of the 38 responses, 7 reported that their state laws and regulations specify a data dictionary with definitions.

Figure 1 below is provided on the all-payer claims database (APCD) Council website. The tool is an interactive map that provides detailed information on the state laws and regulations for APCD.[2]

Figure 1. State laws and regulations for All Payer Claims Databases



References

1. Love D, Rudolph B, and Shah G. Lessons Learned in using Hospital Discharge Data for State and National Initiatives: Implications for CDC's Tracking Program; *J Public Health Management Practice*, 2008, 14(6), 533-542.

2. All Payer Claims Database Council: www.apcdcouncil.org.

Appendices

Appendix A. State Experiences

Missouri Cancer Registry and Research Center (MCR-ARC) Data Linkage Experience

By Saba Yemane

MCR-ARC used the AutoMatch linkage program in the early years of this product's release but started using Registry Plus™ Link Plus in 2006. MCR-ARC uses all Registry Plus™ products and found it very helpful to use Link Plus. MCR-ARC focuses on minimizing both the first type of error, linking non-matches as matches, and the second type of error, rejecting matches as non-matches. As a result, intensive and thorough clerical review needs to be carried out; this is true for any probabilistic linkage program. MCR-ARC has found that Link Plus has very good manual review features that allow for efficient and effective review of potential comparison pairs.

Linkage With Mortality Data: Each year, MCR-ARC links its database with state vital records data to update the vital status of persons already in the MCR-ARC database and to identify deaths attributed to cancer on death certificates for which no corresponding registry case exists (i.e., to identify potentially missed cancer cases). A file is then extracted from the unmatched cases that include any mention of cancer as the underlying cause of death or as any of the contributing factors. The linkage yields, on average, roughly 2,500 potential death certificate only (DCO) cases. Some are followed back to the facility listed as the place of death; missed cases will be abstracted and reported to MCR-ARC. Other potential DCO cases are eliminated through other procedures (e.g., review of the death certificate reveals cause of death was miscoded). About 1,200 cases remain to be linked to the state hospital discharge database, known in Missouri as the Patient Admission System (PAS). database.[1] MCR-ARC creates the file to be linked against the PAS and sends it to the Missouri Department of Health and Senior Services (DHSS). DHSS staff perform the linkage and send MCR-ARC all the information that they have on those cases. MCR-ARC then reviews the cases based on the admission information to determine if the record in the death file is a death clearance case for that year or not and adjusts the abstract accordingly.

After following back to hospitals and linking to PAS, MCR-ARC is generally able to reduce the number of DCO cases to about 650 cases, around 2.2 percent of the total annual incident cases. This process is a collaboration between MCR-ARC and DHSS.

Linkage with Breast and Cervical Cancer Control Data: In Missouri, data linkage with Show Me Healthy Women (SMHW)[2], as Missouri's Breast and Cervical Cancer Control program (BCCCP) funded through CDC's National Breast and Cervical Cancer Early Detection Program (NBCCEDP)[3] is known, is conducted twice per year. SMHW staff send a file with personal identifiers, diagnosis date, and a field specifying the case as breast or cervical cancer to the registry. Cases of breast and cervical cancer are extracted from the MCR-ARC database and prepared for linkage. The SMHW/BCCCP file provided includes approximately 2,000 records; it is linked with a file containing all breast or cervical cancer cases in the registry.

Because Link Plus is programmed for registry operations, it comes with the most up-to-date NAACCR version layout; this facilitates use of Link Plus for these linkages. MCR-ARC creates the record layout for the SMHW/BCCCP linkage file. Linkage is performed and tumor information is provided back to SMHW for matched cases.

This linkage also identifies cases that have not been reported by a reporting facility. Therefore, it is useful as a case-finding tool in identifying potentially missed cases.

MCR-ARC data collection activities are supported by a Cooperative Agreement between the Missouri Department of Health and Senior Services (DHSS) and the CDC and a Surveillance Contract between DHSS and the University of Missouri (#U58/DP003924-04).

References:

1. Missouri Health and Senior Services Patient Abstract System. http://health.mo.gov/data/patientabstractsystem/ accessed August 14, 2015.

2. Missouri Health and Senior Services Show Me Healthy Women. http://health.mo.gov/living/healthcondiseases/chronic/showmehealthywomen accessed August 14, 2015.

3. National Breast and Cervical Cancer Early Detection Program (NBCCEDP). http://www.cdc.gov/cancer/nbccedp/ accessed August 14, 2015.

Pennsylvania Cancer Registry Experience with Link Plus

By Michelle L. Esterly, RHIA, CTR

The Pennsylvania Cancer Registry (PCR) performs probabilistic linkages using Link Plus. Link Plus has been an extremely valuable linkage program and has been utilized for the following linkages conducted by PCR staff for several years:

- De-duplication
- Death Match (state death files and Social Security Death Index [SSDI] files)
- PA HealthyWoman Program[1] – breast and cervical linkage
- Research projects/Confidential data requests
- Reconciliation to verify completeness

Several key features have contributed to the success of using Link Plus for various projects. These include: efficiency, ability to customize configuration by linkage, accuracy of linkage, ability to manually review potential matches with variances in demographic data items, exporting capabilities, and excellent documentation.
Because additional reporting sources often provide unknown or partial SSNs, de-duplication is an important component of quality control procedures to identify duplicate patients within the registry database. Having the ability to modify the cut-off value in Link Plus to accommodate linkages conducted with limited demographic data is essential.

Linkages are conducted on a routine basis between the PCR and the Pennsylvania Vital Records Death files. In addition to obtaining death information to update the cancer registry database, demographic data items such as Race and Hispanic Origin are updated in the database if unknown when known information is available in the Pennsylvania Vital Records Death files.

To prepare for linkage with the National Death Index (NDI), the PCR linked registry data with SSDI files from 1995-2010. Although the manual review process was labor intensive, the benefit to the registry of updating demographic data including SSN prior to linking with the NDI was extremely valuable.

Linkages are also conducted annually between the PA Healthy Woman Program and the PCR to provide additional cancer information to the Breast and Cervical Program.

Link Plus has been utilized for many research projects in Pennsylvania including potential cancer cluster investigations. The registry is interested in promoting use of the data, and Link Plus is one of the tools available to support that effort especially since there is no limitation to file size for linkages.

To verify case completeness from reporting facilities, the PCR conducts a reconciliation process on an annual basis using Link Plus to link facility diagnostic index files to cases received from the reporting source. Listings of potentially missed cases are generated and forwarded to reporting sources requesting reconciliation and submission of missed cases.

References:

1. Pennsylvania Healthy Woman. http://www.portal.state.pa.us/portal/server.pt/community/healthy_women/14172/healthywoman_program_home/557855 accessed August 14, 2015.

The Illinois State Cancer Registry Data Linkage Experience

By Lori Koch, BA, CCRP, CTR

The Illinois State Cancer Registry (ISCR) performs both deterministic and probabilistic linkages utilizing various software programs including Link Plus and Microsoft Access®. These linkages help the registry de-duplicate cases, share information between cancer control programs, provide data to researchers, and maintain high quality data by updating and adding information.

As do many registries, ISCR links its registry database to external sources to update or improve various data items. Each year, ISCR performs a linkage between the Illinois Vital Records Death Master File and the ISCR registry database. This project allows ISCR to obtain specific data items currently coded as unknown or left blank on the ISCR registry database. Items such as maiden name, race, Hispanic ethnicity, and gender are obtained through this linkage. Additionally, ISCR obtains updated death information including vital status, cause of death, and date of death. Attempts are also made to update unknown items like SSN through linkage with either the Illinois Vital Records Death Master File or SSDI. Having known and accurate SSNs can be an important component in achieving

high quality linkages; however, in recent years, the loss of the SSDI as a resource has made this more difficult.

ISCR also uses linkage to determine which cases are not in the registry database but have a cancer cause of death recorded on the death certificate. These cases must be followed back during the death clearance process each year. A combination of deterministic and probabilistic linkages is conducted, both for person-to-person linkage and tumor-to-tumor linkage. ISCR also uses linkage to determine alternative or current address and physician information for use in the death clearance process. ISCR obtains yearly files from both Reference USA[1] and the Illinois Department of Professional Regulation[2] and then links those files by physician name, specialty, and city to determine address information.

ISCR uses the Link Plus program to identify duplicate reports in its registry database as a part of ISCR's quality control program. The Rocky Mountain Cancer Data Systems (RMCDS)[3] software used by ISCR allows the user to perform a deterministic record linkage using selected match passes in order to identify possible duplicates. The Link Plus program allows the user to perform a probabilistic record linkage using manually determined match passes and weights which also helps to identify possible duplicate cases. Both types of linkages are performed at the end of the month on all cases that have been reported by multiple reporting facilities within that month. Personnel must then determine which cases to remove from the central registry database by manually reviewing possible duplicates.

The Illinois Breast and Cervical Cancer Program (IBCCP)[4] and the ISCR registry database are linked annually using Link Plus to provide staging and tumor size information while also providing confirmation of diagnosis. This linkage gives ISCR an opportunity to identify unreported cases and to obtain previously unknown information on race and ethnicity.

Providing cancer data to researchers is an important component of ISCR's commitment to fully utilizing the cancer registry data we collect. ISCR participates in a wide range of studies on various types of cancer with researchers located both in and out of the State of Illinois. Because of this high level of participation in research, ISCR staff have become adept at conducting large file linkages using both probabilistic and deterministic methods and various types of software. These linkages allow researchers to determine cancer diagnoses, treatment, and outcomes.

References:

1. Reference USA. http://www.referenceusa.com/ accessed August 14, 2015.

2. State of Illinois Department of Financial and Professional Regulation. http://www.idfpr.com/ accessed August 24, 2015.

3. Rocky Mountain Data Systems. http://rmcds1.med.utah.edu/ accessed August 24, 2015.

4. Illinois Breast & Cervical Cancer Program. http://cancerscreening.illinois.gov/ accessed August 14, 2015.

Hospital Discharge Linkage Methods and Results from Massachusetts Cancer Registry

By Richard Knowlton, MS

Application for the Use of Hospital Discharge Data: As a component of the Comparative Effectiveness Research (CER) Payer Project[1], the Massachusetts Cancer Registry (MCR) compared insurance data from Massachusetts hospital discharge data with both insurance data originally reported to the MCR and insurance data collected during medical record reviews. In February 2012, the MCR submitted an application to the Massachusetts Department of Public Health's Division of Health Care Finance to obtain inpatient hospital discharge data from 2005-2009. The MCR made the decision to link only inpatient data based on the very low yield from linking emergency room and outpatient data during the previous primary payer project and the relatively high yield from inpatient data. The application was approved at the end of April 2012. Name and SSN were not available on the hospital discharge data. The best identifying data available were medical record number, hospital code, and date of birth.

File Creation: The hospital discharge data (HDD) were made available to the MCR through a SQL server. The MCR epidemiologist wrote a SAS program to access the following HDD variables from the server: hospital code, date of admission, date of discharge, medical record number, date of birth, 15 diagnosis variables, 15 surgical variables, primary insurance, primary insurance type, secondary insurance, secondary insurance type, Zip code, sex, and unique ID number from the HDD. The HDD data are released by fiscal year (FY), meaning that FY05 covers hospitalizations from July 1, 2004, to June 30, 2005. This being the case, the MCR requested HDD FY05 to HDD FY10 to cover the diagnosis years 2005-2009 from the MCR/CER sampled data. As a result, six SAS programs exported six HDD text files with the previously listed variables. The MCR also created a text file of the 5,000 MCR cases selected for the CER project. This file contained the following variables: MCR/CER ID number, hospital code, birth date, medical record number, date of diagnosis, date of first surgery, date of most recent surgery, date of radiation, date of chemotherapy, and date of hormone treatment.

Linkage: The MCR used Link Plus, a database linkage program created by CDC's National Program of Cancer Registries (NPCR), to match the HDD and MCR/CER files. The entire MCR/CER file was matched six times against each year of HDD. Although individual years of the MCR/CER file could have been matched against individual years of the HDD, the MCR decided to match the entire MCR/CER file against each HDD year to account for overlaps in data (i.e., diagnosed in one fiscal year and treated in another). Link Plus assigned scores based on how strong a data element matched. Matches were evaluated by this score and by reviewing borderline matches and checking for birth dates or medical record numbers that were slightly off and determining if there was a genuine match. In many instances, a medical record number in the MCR/CER file was missing an alpha prefix which was in the HDD file. If the numbers matched in both files despite the alpha prefix, it was considered a match.

Results: Of the 5,348 cases in the MCR/CER sample database, there were 2,029 cases (38%) that matched with a record in the HDD based on medical record, hospital code, and date of birth. A majority of the cases matched only on dates of treatment and not date of diagnosis. Because the HDD reliability analysis for this study involved only a match on payer at diagnosis, only 715 (35%) of these matches met the criteria and were included in the analysis. The weighted results of the

agreement analyses between payer data from the inpatient hospital discharge database and the re-abstracted payer at diagnosis are presented in Tables 1 and 2, one for the pre-reform period (2005-2006) and one for the post-reform period (2007-2009). The numbers presented are weighted. Compared to the overall sampled cases, the matched HDD cases were overly representative of colorectal cancer cases. This is likely due to the HDD being inpatient data and more colorectal cancer patients being diagnosed while in the hospital. For both time periods, the agreement was very strong for private insurance and overall Medicaid and Medicare. The agreement was less strong for subcategories of these two insurance groups, but that may be a reflection of how Medicare and Medicaid subgroups are presented in the hospital discharge data. The agreement for Commonwealth Care, the healthcare reform insurance, in the second time period was 100 percent. Agreements for free care and no insurance were not as strong.

Table 1. Agreement between hospital discharge payer at diagnosis and re-abstracted payer at diagnosis, weighted sample, 2005-2006

Insurance in Discharge Data	Total (n = 2,811)		Breast (n = 391)		Colorectal (n = 2,419)	
	% Agreement	Kappa Score	% Agreement	k (CI)	% Agreement	Kappa Score
No Insurance	50.7	0.50	100.0	1.00	33.3	0.33
Free Care, Health Safety Net	24.6	0.17	.	.	33.4	0.19
Private: Managed Care/FFS	91.5	0.89	90.0	0.80	91.8	0.90
Medicaid:	64.7	0.78	66.7	0.79	64.3	0.77
Medicaid Alone	59.9	0.71	50.0	0.47	61.5	0.75
Medicaid Managed Care	48.5	0.65	.	.	100.0	1.00
Medicare:	94.3	0.86	95.5	0.95	94.1	0.85
Medicare Alone	51.2	0.47	80.0	0.60	47.7	0.45
Medicare Managed Care	77.9	0.75	.	.	85.7	0.80
Medicare With Supplement	83.0	0.72	58.1	0.54	85.6	0.74
Medicare With Medicaid	74.4	0.75	100.0	0.84	71.9	0.74

Table 2. Agreement between hospital discharge payer at diagnosis and re-abstracted payer at diagnosis, weighted sample, 2007-2009

Insurance in Discharge Data	Total (n = 2,811)		Breast (n = 391)		Colorectal (n = 2,419)	
	% Agreement	Kappa Score	% Agreement	k (CI)	% Agreement	Kappa Score
No Insurance	50.1	0.50	.	.	50.1	0.50
Free Care, Health Safety Net	32.7	0.28	.	.	50.1	0.33
Commonwealth Care	100.0	1.00	100.0	1.00	100.0	1.00
Private: Managed Care/FFS	86.4	0.86	100.0	0.94	84.9	0.85
Medicaid:	84.4	0.86	100.0	1.00	78.5	0.81
Medicaid Alone	86.8	0.89	100.0	1.00	83.3	0.87
Medicaid Managed Care	75.7	0.76	100.0	1.00	50.0	0.50
Medicare:	94.6	0.83	94.3	0.91	94.7	0.82
Medicare Alone	33.1	0.27	16.6	0.08	36.7	0.31
Medicare Managed Care	84.8	0.78	100.0	1.00	81.8	0.74
Medicare With Supplement	83.2	0.75	100.0	0.84	81.6	0.73
Medicare With Medicaid	69.9	0.69	50.0	0.56	77.8	0.73

The results of the hospital discharge match indicated a strong agreement between re-abstracted payer at diagnosis and payer at diagnosis in the hospital discharge data for private insurance, Medicaid, and Medicare for the 2005-2009 period. The agreement for Commonwealth Care (100%) was also very strong for the 2007-2009 period.

References:

1. Comparative Effectiveness Research Data Collection Enhancement Project. http://www.cdc.gov/cancer/npcr/cer_data_collection.htm accessed 8/25/2015.

Nebraska Cancer Registry Hospital Discharge Data Linkage

By Ge Lin, PhD

This section describes linkage methods and some results for linking Nebraska Cancer Registry (NCR) data with Nebraska hospital discharge data (NHDD). In September 2011, the registry requested NHDD data for 2005 to 2009, which allowed time for delayed reporting and ensured fairly complete case ascertainment for those years. The identifiable data items in the NCR are patient ID, social security number, first name, middle name, last name, date of birth, sex, street address at diagnosis, and treatment facility ID. The NHDD maintained by the Nebraska Hospital Association has information about inpatient, outpatient, and emergency room (ER) services. Pertinent patient-identifiable data elements in the NHDD include all those in the NCR except SSN. Although the date of admission is useful for some cancer patients in the NCR file when the dates of the first course of treatment are available, they are not used for linkage. For the 5-year study period, the NHDD had more than 13,041,435 hospital records from Nebraska's 87 non-military hospitals: 1,052,493 for inpatient, 2,104,917 for ER outpatient, and 9,884,035 non-ER outpatient services.

Treatment and billing information in the NHDD are based on the standard UB-04 form. Helpful inpatient data items for a cancer registry include diagnostic codes and procedure codes for surgery. The former can be used to construct a comorbidity index and the latter can be used to verify some surgical procedures for the cancer registry. Helpful outpatient information includes Current Procedural Terminology (CPT) codes, which are used primarily for insurance billing purposes, with each treatment being assigned a corresponding CPT code.[1] Because non-surgery treatment information is generally not very good in a cancer registry and chart review is very expensive, using CPT codes to update some treatment information fits nicely to any project linking a cancer registry with a hospital discharge system. For instance, one can use CPT codes for colonoscopies to check date of last screening. However, one has to be knowledgeable about both inpatient and outpatient procedure codes. For example, there are roughly 12 CPT codes for colonoscopies, but not all of them are for cancer screening. Another obstacle to using CPT and ICD-9[2] treatment codes is that a filing for Medicaid may bundle several related treatments into one code.

Due to the lack of a unique identifier (e.g., social security number) between the two datasets, Link Plus 2.0 was used for data linkage.[3] (Link Plus 2.1 can be used for linking two datasets or de-duplicating one dataset.) The linkage process was designed according to Newcombe's four steps[4]: (1) data preparation, (2) matching and merging, (3) manual review, and (4) verification.

Data Preparation: Data preparation included checking data quality, de-duplicating, parsing, and standardizing the linkage variables common in both datasets. The process started with the NCR data for January 1, 2005, to December 31, 2009, with 54,990 records, which included both in-state and out-of-state patients. The de-duplication process was based on all linkage variables: first name, last name, date of birth, sex, county and Zip code of residence, and primary cancer site. This step resulted in 52,027 unique patients.

Because the 5-year NHDD file was extremely large (more than 10 million records), it was divided into cancer-related and non-cancer-related datasets to increase computational efficiency. For each record, ICD-9-CM diagnostic codes for up to 10 diagnoses were searched.[5] Those in the range of 140 to 208 or equal to 2386 were classified as cancer related.

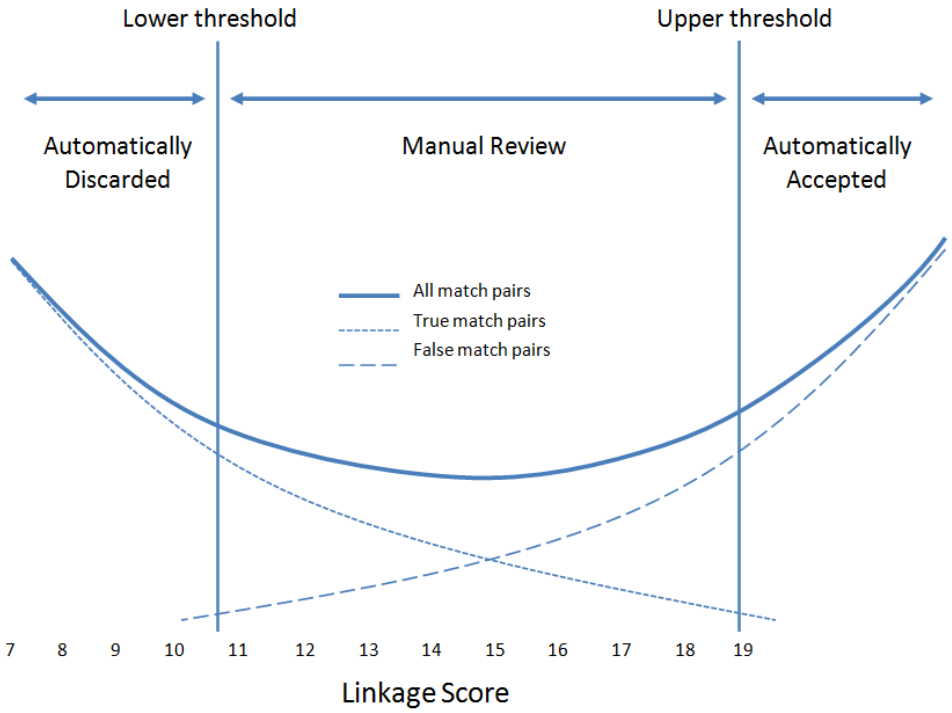
Matching: The NHDD was divided into two datasets, one for cancer-related admissions and one for all other admissions, which in turn allows for the blocking design. When a linkage is complete, Link Plus calculates a score (total linkage weight) for all comparisons of a case in the NCR file with a case in the NHDD file and generates a default cut-off weight above which values are considered as potential matches.[6] Figure 1 provides a diagram showing the relationship between linkage scores and linking decisions. Low linkage scores, which indicate a non-match, start from the left and the score increases along the X-axis toward the right. At the center point (e.g., 10), the score suggests a potential match; manual review is needed. As the score increases to between 16 and 17, it starts to suggest true matches for the remaining pairs and manual review is no longer required.

Manual Review: All potential matched pairs above the cut-off weight were reviewed manually after each linkage run. A pair was considered as a true match if there was a minor discrepancy, such as an obvious typographic mistake, in the matching variables: transposed first and last names; either the first name, the last name, or the middle initial matched with another unmatched name; variants of a first name that the Soundex did not identify; a hyphenated first name or last name; transposed birth

month and day; and closeness of the date of cancer diagnosis and the date of hospital admission. After manual review, 53,301 NCR records corresponded to 1,102,826 NHDD records.

Validation: To validate the potential match of 54,990 records, a random sample calculation can be used to decide how many pairs need to be checked independently for validation. In this case a random sample of 381 records would be sufficient at the 95 percent confidence level. Therefore, 400 linkage pairs were randomly selected for an independent manual review. Three trained data analysts performed the manual evaluation and found very few false positives (3, 1, and 0 pairs as false positives, respectively, by the three reviewers).

Figure 1. Two-threshold scheme for linkage scores using manual review



Linkage Quality: As mentioned above, NCR was able to link 97 percent of its records (53,301 out of 54,990 records). This is considered high because some records extracted from autopsy and death records may not even be in an area hospital. For those records that indicated a hospital procedure, NCR found 99.99 percent of them if the hospital was relatively large (more than 50 beds).

NCR found about 1.2 million NHDD records for 50,426 unique and linked NCR patients suggesting that each patient had more than 20 inpatient and/or outpatient hospital visits. Using CPT codes for radiation treatment (RT), NCR was able to increase RT rates by 10 percent to 15 percent, depending on the cancer site. Finally, NCR could identify a number of comorbidities for each cancer case.

References:

1. Abraham M, Ahlman JT and Boudreau A, et al. CPT 2011: standard edition. Chicago, IL: American Medical Association Press.

2. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (http://www.cdc.gov/nchs/icd/icd9cm.htm). Accessed May 17, 2012.

3. CDC 2006. Link Plus Version 2.10 probabilistic record linkage software. Atlanta, GA: Centers for Disease Control and Prevention, 2006.

4. Newcombe HB. Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford, England: Oxford University Press, 1988.

5. Penberthy L, McClish D and Pugh A, et al. Using hospital discharge files to enhance cancer surveillance. *Am J Epidemiol* 2003; 158: 27-34.

6. Blakely T and Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol* 2002; 31: 1246–1252.

Usefulness of Billing Data from Community Specialty Practices for Supplementing Treatment Reporting to Central Registries

By Lynne Penberthy MD, MPH

Challenges to Cancer Surveillance: Gaps in reporting of initial course of therapy for cancer surveillance result in limitations to the usefulness of the cancer surveillance data maintained by CCRs, including the NCI SEER and CDC NPCR registries. The data are incomplete for several reasons. First, because the primary source for case and treatment reporting is hospital registries.[1-5] Hospital-based registrars lack access to information on treatments provided to patients outside the hospital system such as in community practices. This lack of access, coupled with a simultaneous migration of systemic treatments for cancer to the outpatient community provider setting,[6] has further increased the lack of treatment information in registries. These community providers are less likely to report to the CCR, even when mandated to do so by state statutes, laws, and regulations. Finally, diminishing resources limit the capacity of the CCR to perform active surveillance and/or active follow-back. All these factors have resulted in low sensitivity of cancer registries to systemic therapy and the potential for underreporting as well as biased and/or delayed reporting of treatment.

Why is Capturing Treatment Important? The biases inherent in limiting registry information to treatment administered in hospitals may misrepresent true patterns of care. With incomplete treatment reporting, registries are unable to provide information on outcomes in the context of treatment for the non-clinical trial population (representing > 95% of cancer patients).

Mortality is not the only outcome of importance. There are more than 14 million current cancer survivors in the United States. Lack of complete treatment for those patients prevents both an understanding of the long-term effects of treatments and the identification of subsequent treatments for recurrent disease.

Therefore, capturing treatment is critical to understanding cancer care and outcomes. Traditional registry methods including manual abstraction and active follow-back are no longer viable solutions. Alternative methods need to be developed and tested; in particular, methods for automated capture of this information are critical given the decline in resources.

One such alternative is using commonly available and standardized billing data. Capturing these data directly from the practice will include claims for all insurers and patients of all ages. The billing data in the standardized 837 Professional format include standardized nomenclature (ICD-9 diagnosis codes and CPT codes) as well as patient demographic data and physician information. Claims data have been demonstrated to have high reliability and accuracy for reporting detailed treatment information. Registries can focus on high-yield practices (such as oncology, dermatology, or urology) to maximize value. Once such a system is implemented, it requires minimal maintenance and automatically reports data on all treatment as well as providing follow-up information on the patients. A software system has been developed and tested in seven oncology practices, a large urology practice and a large HMO/PPO to evaluate the ability of the system to identify cases and to supplement treatment information on known cases. This brief summary provides results on the validity and accuracy of the codes and briefly describes how the system works and its availability.

Methods: *Automated software* - the software was developed under a contract and grant from the National Cancer Institute (NCI) and is freely available. It is written in C#, (C-Sharp, an object-oriented programming language in the C family) on a .NET framework. All data are stored in a SQL Server database.

It receives submitted billing data in a standardized format (837 Professional-4010 or -5010 format). Billing includes codified data elements that represent diagnosis (ICD-9 Codes) and detailed information on treatment Healthcare Common Procedure Coding System (HCPCS) codes. The latter have been demonstrated to have high validity, sensitivity and specificity.[2-4, 7-8] The software automatically screens the billing data for cancer diagnoses and treatments. Once a patient is identified as having a cancer diagnosis, all subsequent treatment information is captured and stored in the database. ICD-9 codes are mapped to ICD-O-3 cancer codes, and the latter are used to populate the diagnosis field in the NAACCR record that will be subsequently generated. Treatment data are also captured and automatically mapped to Facility Oncology Registry Data Standards (FORDS) codes where appropriate. However, more detailed information (e.g., chemotherapeutic agent, dates administered, etc.) is maintained in linked treatment tables. These data are used to populate a SQL Server database and tables specific to: demographics; diagnosis (including probable dates of diagnosis based on first occurrence in billing); comorbidity; and specific treatment tables for each of the categories of treatment, including surgery, chemotherapy, radiation therapy, hormonal therapy, and biologic response modifiers. The software then combines data from the tables to automatically generate a partial NAACCR record to send to the CCR at scheduled intervals. The default interval is 1 year but can be redefined by the user. The partial NAACCR record includes 56 fields. The system is available on request from the author but will require technical expertise to install and set up to run either at the practice location or at the central registry.

Pilot Testing: The system has been piloted in a variety of clinical settings: a large urology practice in New Jersey,[9] six community oncology practices in North Carolina and Virginia, and a large HMO/PPO in Los Angeles. Detailed results from these pilot studies are available in prior publications.[9,10] The system is currently in testing on a larger scale in the Florida Cancer Data

System (FCDS), where it is being adapted for receiving data from multiple practices at a central location.

Results: A summary of the pilot studies, including the study interval, participating providers, number of cancer patients, and basic treatment captured, is provided below in Table 3. The number of providers in each of the pilots ranged from hundreds (Los Angeles SEER pilot) to 29 in the NC pilot study. The number of claims processed reflected both the number of providers and the study interval, and ranged from a low of 19,700 to nearly 1 million. In the pilot studies, any patient with a single cancer diagnosis was included in the database. This impacted the solid tumor match rate shown in the table, which ranged from 71 percent to 87 percent.

Table 3. Summary of four pilot projects to evaluate the ability of billing data from community specialty practices to supplement treatment for incident matched cases

Project	Study Interval Months	N Providers	Number Claims Processed	Number Cancer Patients	% Match Rate	Number Treatments Identified	% Patients With Billing RX
LA SEER HMO/PPO	6	26 Spec	923,000	11,863	70.9	32,680	24.0
NC Oncology Practices	8 + 12	29	19,717	11,535	75.0	40,963	23.7
NJ Urology Practice*	3.5	35	26,000	2,170	87.2	1,256	57.9
FL Cancer Data System	ongoing	>500	500,000	?	100%**	>100,000	100%**

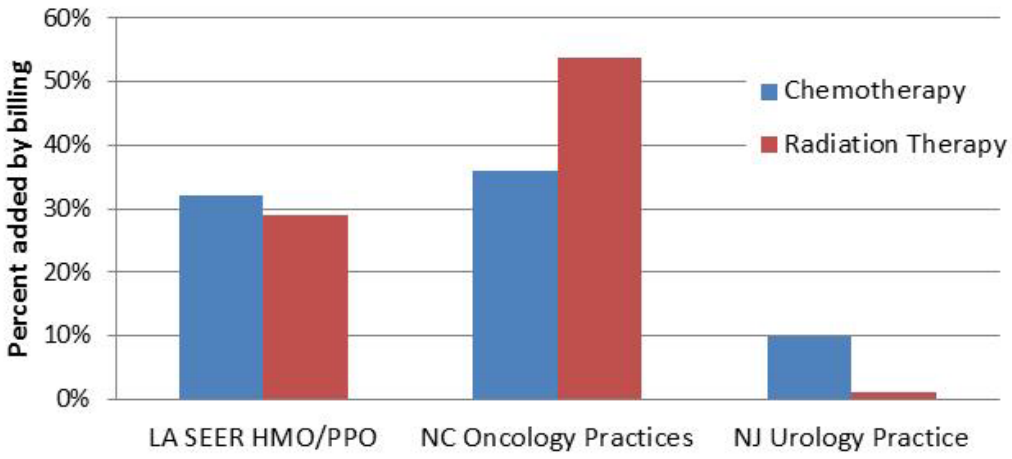
* Note: Urology practice was manually reporting cancer cases to the CCR during the study period.
** Currently, Florida is testing only for supplementing RX for known cases.

The accuracy of the ICD-9 codes for identifying cancer cases, despite this very sensitive case definition, was good to excellent, ranging from 95 percent overall for both solid tumors and hematologic malignancies from oncology practices to 96 percent for urologic malignancies captured from the large urologic practice. Based on review of incorrectly captured cases for the urologic malignancies, it was found that increasing the requirement for case finding to having >1 diagnosis over several months increased the accuracy of the billing to 99.5 percent.

The accuracy of treatment data was very good. For 566 patients in Virginia, 100 percent of billing-reported treatments were validated from the medical record. For the urologic practice, only one treatment of the 152 reported in billing could not be verified in the medical record, representing a 99.3 percent accuracy rate.

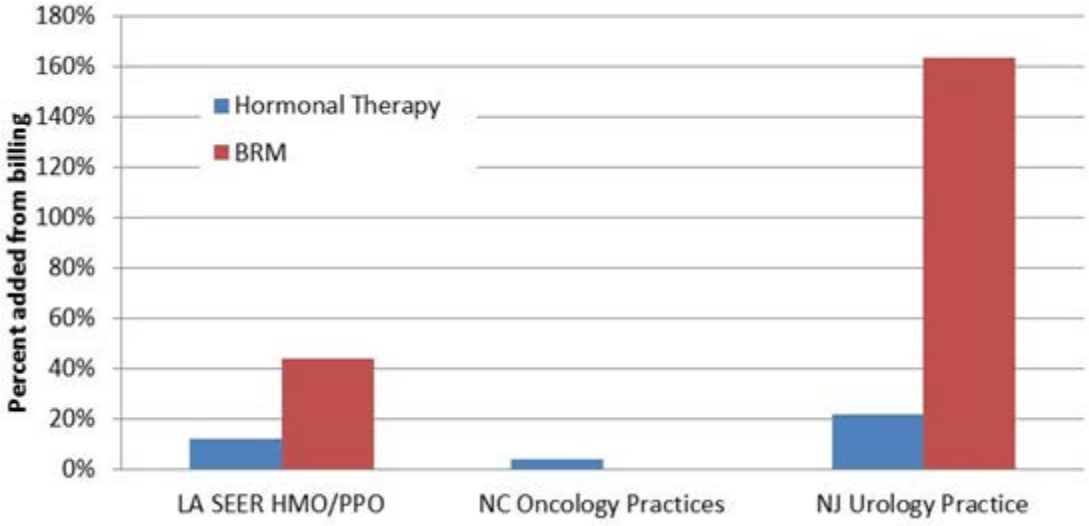
Given the very high rates of accuracy, the ability of the billing information to contribute additional information to incident cancers already reported to the central registry was evaluated. Figures 2 and 3 represent a summary of how much billing data can contribute to previously reported incident cancers. For chemotherapy and radiation, the billing data from oncology practices provided information on an additional 32 – 38 percent of patients for chemotherapy and 39 -54 percent of patients for radiation. For urology it was lower, only 10 and 1 percent respectively due to treatment patterns for the key urologic malignancies (prostate and bladder) and the very recent addition of a radiation machine at the practice (during the pilot study).

Figure 2. Percent of chemotherapy and radiation therapy contributed by billing data among matched incident solid tumors



For hormonal therapy and biologic response modifiers, the results were slightly different, with only a small number of billing reports providing additional information on hormonal therapy (10 – 22 percent), but 43 – 163 percent added treatment encounters were recorded for biologic response modifiers as shown below.

Figure 3. Percent of hormonal therapy and BRM contributed by billing data in matched incident solid tumors



In addition to the information available on known incident cancers, the billing data automatically provide information on subsequent courses of therapy. There were significant additions to the number of treatments subsequent to initial therapy identified for prevalent cancers in the Los Angeles SEER registry when matched with 6 months of billing data from the large HMO/PPO included in that pilot study. The percentages of added treatments were 13 percent for radiation, 17

percent for chemotherapy, and 33 percent for hormonal and 34 percent for biologic response modifiers.

Limitations: There are several limitations to the system and to billing data by themselves. The system currently attributes systemic treatment to more than one cancer when more than one cancer diagnosis code occurs on the bill. This occurred in 3 – 7 percent of patients in the studies. These were reviewed and approximately 85 percent of the treatments were assigned to the appropriate cancer site. Nevertheless, some records will require manual review and potentially follow-back to the physician practice. Also, systemic treatment that is provided as an inpatient (e.g., chemotherapy) may be missed using billing data. Inpatient billing records do not allow for the HCPC codes; rather, they report ICD-9 procedure codes, which do not represent specific agents. Therefore, while the billing data from community specialty practices are highly accurate, they are not always complete. Billing data do not capture orally administered treatments (with a few exceptions where there is an equivalent oral and intravenously administered agent). Finally, the billing data do not include stage or histology (except for hematologic malignancies). Linking billing codes with text-based information that might be reported in the electronic health record under Meaningful Use (MU) criteria might serve as an efficient mechanism to supplement billing-reported information that may be available only in the electronic medical record (e.g. stage). The merger of the two could expand the amount of discrete data captured, reducing the manual review of the data submitted under the MU requirements for cancer reporting.

Summary: Using billing data from community oncology practices to supplement both casefinding and treatment reporting is likely to provide high-quality and useful data to registries. The information provided may be substantial, and automation offers an efficient method of quickly and consistently screening, capturing, and reporting information on cases and/or treatment that have a high probability of being missed. Using such tools to supplement more traditional methods will likely begin to fill the gaps in treatment data for cancer surveillance.

Acknowledgements:

Practices:

- Dr. Gregory Formanek Virginia Physicians Inc., Richmond, VA
- Hematology Oncology Patient Enterprises PC, Charlottesville, VA
- US Oncology - Cancer Centers of North Carolina, Raleigh and Asheville, NC
- Health Care Partners Los Angeles CA
- Delaware Valley Urology, Voorhees NJ

Registries:

- The Virginia Cancer Registry
- The North Carolina Central Cancer Registry
- The California Cancer Registry
- The New Jersey Cancer Registry
- The Florida Cancer Data System

Co-investigators/Colleagues:

- Laurel Gray
- Jim Martin
- Lynn Sribor
- Dennis Deapen
- Soundarya Radhakrishnan
- Sandra Overton
- Pam Agovino
- Donna McClish
- Valentina Petkov
- Chris Gillam
- Davis Gentry

This work was supported by the National Cancer Institute at the National Institutes of Health (grant number R21 CA127967-01) and the National Cancer Institute/ Information Management Systems, Inc. (Subcontract D5-VCU-1) and RFP NO1 PC 95002-18 Statement of Work 09-7.

For additional information on accessing the software, please contact:
Lynne Penberthy MD, MPH
E-mail: lynnepenberthy.schumacher-penberthy@nih.gov
Phone: 240-276-6864

References:

1. Lund JL, Sturmer T, Harlan LC, et al. (2013) Identifying specific chemotherapeutic agents in medicare data: a validation study. *Med Care* 51(5): e27-e34.

2. Du XL, Key CR, Dickie L, Darling R, Geraci JM, Zhang D (2006) External validation of medicare claims for breast cancer chemotherapy compared with medical chart reviews. *Med Care* 44(2): 124-131.

3. Lamont EB, Lauderdale DS, Schilsky RL, Christakis NA (2002) Construct validity of medicare chemotherapy claims: the case of 5FU. *Med Care* 40(3): 201-211.

4. Lamont EB, Herndon JE, Weeks JC, et al. (2005) Criterion validity of Medicare chemotherapy claims in Cancer and Leukemia Group B breast and lung cancer trial participants. *J Natl Cancer Inst* 97(14): 1080-1083.

5. Warren JL, Harlan LC, Fahey A, et al. (2002) Utility of the SEER-Medicare Data to Identify Chemotherapy Use. *Medical Care* 40(8): IV55-IV61.

6. Decker SL, Schappert SM, Sisk JE (2009) Use of medical care for chronic conditions. *Health Aff* (Millwood) 28(1): 26-35.

7. Penberthy L, McClish D, Manning C, Retchin S, Smith T (2005) The added value of claims for cancer surveillance: results of varying case definitions. *Med Care* 43(7): 705-712.

8. McClish DK, Penberthy L, Whittemore M, et al. (1997) Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol* 145(3): 227-233.

9. Penberthy LT, McClish D, Agovino P (2010) Impact of automated data collection from urology offices: improving incidence and treatment reporting in urologic cancers. *J Registry Manag* 37(4): 141-147.

10. Penberthy L, McClish D, Peace S, et al. (2012) Hematologic malignancies: an opportunity to fill a gap in cancer surveillance. *Cancer Causes Control* 23(8): 1253-1264.

Probabilistic Linkage Methodology for Match of Breast Cancer Cases, July 2009

By Therese A. Dolecek, PhD and Daniel R. Leonard, MS

The following method was used to probabilistically link records from the Illinois State Cancer Registry (ISCR) to records from the Illinois Hospital Discharge Data (IHDD) for inpatient and outpatient surgery admissions. A total of 29,381 ISCR breast cancer case records matched in a one-to-many relationship to 44,696 hospital records for diagnosis years 2002 through 2005.

Data Sources and Record Selection:

A. ISCR

The Illinois Department of Public Health (IDPH) Division of Epidemiologic Studies provided 38,247 ISCR records with a primary site of breast cancer diagnosed between 1/1/2002 and 12/31/2005. The records were restricted to females.

B. IHDD

The IDPH Division of Patient Safety and Quality provided 2,961,063 inpatient and 2,961,685 outpatient surgery records for patients with discharge dates between 1/1/2002 and 12/31/2005. The cases were also restricted to females.

Selection of Subset of Breast Cancer Cases: Breast cancer cases were selected in the IHDD records using criteria from the Agency for Health Care Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP) (see Appendix B for details). This resulted in a subset of 161,741 candidate IHDD linkage records for women who received inpatient and/or outpatient surgery.

It should be noted that these criteria also include records with only a history of breast cancer (ICD-9 code V10.3) as well as current diagnoses. History of diagnosis records are much less likely to link than current diagnoses, but they were maintained for the sake of consistency as well as to provide background cases for evaluating the linkage results.

Identification of Related Procedures: To further help with evaluating the linkage, procedures associated with breast cancer were flagged. These consisted of mastectomy, lumpectomy, and biopsy again identified principally using AHRQ criteria (see Appendix C for details). The three types of procedures were flagged separately because the same record could contain more than one surgery code. There were 54,078 patients with at least one such procedure in the subset of 161,741 IHDD records.

Linkage Algorithm and Procedures: The linkage was performed using AutoMatch, an older but highly flexible linkage software developed by MatchWare Technologies, Inc. To help describe the algorithm, a section of the actual code is shown in Figure 4. The full code is provided in Appendix D.

Figure 4. Coding from the linkage algorithm

```
PROGRAM GEOMATCH MULTIPLE
DICTA hosp
DICTB iscr
;
;
BLOCK1 CHAR hosp hosp
BLOCK1 CHAR ZIP ZIP
BLOCK1 CHAR DOB DOB
;
MATCH1 CHAR hosp hosp 0.89 0.01
MATCH1 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH1 DATE8 admdate datedx 0.99 0.1 3 90
MATCH1 CNT_DIFF zip zip 0.99 0.01 1
;
CUTOFF1 23.9 23.7 23.9
;
;
BLOCK2 CHAR hosp hosp
```

A. PROGRAM GEOMATCH MULTIPLE

This linkage was designed as a one-to-many match. It allows for the possibility that one ISCR record could match to multiple hospital admissions for either inpatient or outpatient surgery. Multiple admissions were considered a likely experience for many patients.

1. DICTA hosp
   DICTB iscr
   These statements simply identify the text files with the actual data.
   (hospital discharge record file (DICTA hosp);
   ISCR breast cancer case record file (DICTA iscr)

2. BLOCK1 CHAR hosp hosp
   BLOCK1 CHAR ZIP ZIP
   BLOCK1 CHAR DOB DOB

AutoMatch and other linkage software programs use “blocking.” Rather than compare every record in the first file to every record in the second file at one time, restricted subsets or “blocks” are made within which records are compared. The full match uses several different blocks, with a “pass” through the entire dataset for each.

As commonly performed, the first block (BLOCK1) is very restrictive, nearly forcing an exact match. Here block 1 requires that records from the two files be identical on hospital name (hosp), zip code (ZIP), and date of birth (DOB) even before comparing them to calculate match weights.

Note that, later in Figure 4, BLOCK2 uses only hospital name (Appendix D). A common practice after the first block is to use one blocking variable at a time. In this manner, a true match between records that had a slightly different ZIP code or DOB would be captured in the second pass.

3. MATCH1 CHAR hosp hosp 0.89 0.01
   MATCH1 CNT_DIFF DOB DOB 0.99 0.01 1
   MATCH1 DATE8 admdate datedx 0.99 0.1 3 90
   MATCH1 CNT_DIFF zip zip 0.99 0.01 1

This coding lists the variables for which actual weights will be calculated. For this linkage, matching variables consisted of hospital name (hosp), date of birth (DOB), date of diagnosis (datedx) relative to date of admission (admdate), and patient ZIP code of residence (zip).

AutoMatch offers great flexibility in the types of variables it can work with and how they are compared. Here for hospital names (hosp), the match needed to be exact. For DOB and ZIP up to 1 digit could differ.

For dates, the parameters were set so that a full weight for matching was given if the date of diagnosis was exactly the same as the admission date. A full negative weight against matching was given if the admission date came 90 days or more after the diagnosis date. Between these two dates, the weight was prorated. A similar prorating took place for admission dates preceding diagnosis dates by up to 3 days. In practice, based on the resulting weight values, this forced matches into a narrow window. They only included records with an admission date up to 2 days before the diagnosis date and up to 60 days after it.

The numbers at the end of each MATCH line help define how to work with each variable (as just described) and also provide the *m* and *u* values associated with probabilistic linkage theory. The *m* value basically concerns the reliability of the variable, and the *u* value basically concerns the chances of a random match on the variable. In AutoMatch, the actual *m* and *u* values are calculated during the match process and overwrite the placeholder values in the above coding.

It should be noted that the IHDD ICD-9 diagnosis codes were not used as a matching variable. In discussion with IDPH staff, it was thought that these codes might be inconsistent with the ISCR ICD-O-3 primary site codes between the two data sets. Also the match worked very well without this variable so it was not necessary.

4. CUTOFF1 23.9 23.7 23.9

This line of coding is an important control step to define cutoff weights for the match. These weights will be described in detail below in Section B.

5. BLOCK2 CHAR hosp hosp

As noted above, a separate block was set up using hospital names (hosp). Five blocks were created and so five passes were used. Blocks 3 and 4 consisted of, respectively, DOB and ZIP. The fifth

block consisted of admission date equal to diagnosis date. For each block the matching variables were the same. (See Appendix D with complete linkage algorithm.)

B. CUTOFF WEIGHTS

Cutoff weights are manually entered in AutoMatch for each block (i.e., each "pass" through the data). A very useful approach to working with these weights was developed in association with National Highway Traffic Safety Administration linkage projects.[1,2] For this approach, two sets of "odds" need to be defined: Prior Odds and Posterior Odds.

1. Prior Odds (PrO)

PrO is defined as the ratio of true match comparisons to non-match comparisons in the linkage. In this case, every record in each dataset is a linkage candidate (blocking does not apply here).

PrO is calculated using a few simple steps. First, the total numbers of records for each data set are multiplied to obtain the total number of possible comparisons. The ISCR file contained 38,426 records and the IHDD file contained 161,741 records. Multiplying these gives a total number of 6,215,059,666 possible comparisons.

Second, the number of match and non-match comparisons is calculated. This linkage had 44,696 matches. By subtraction, this leaves 6,215,014,970 non-matches.

Finally, PrO is calculated as the ratio of match to non-match comparisons, in this case 44,696 / 6,215,014,970. This is approximately 1 in 139,051. Essentially, these are the "odds" that the match weights need to overcome to separate matches from non-matches.

2. Posterior Odds (PoO)

PoO is defined as the desirable separation between matches and non-matches. PoO does not need to be calculated. It is simply a ratio, with typical values of 9-to-1, 99-to-1, etc.

PrO and PoO can be combined and then compared with the actual match weights to develop a cutoff value. For example, dividing the PoO of 99-to-1 by the PrO gives a value of 13,766,030. In Log Base 2, this becomes 23.7, and this value can be compared with the weights produced by AutoMatch.

AutoMatch calculates weights using the previously noted *m* and *u* values for matching variables. A simple way to think about this calculation is to consider the random chance of matching on both ZIP code and hospital name. If the random chance of each of these occurrences is 1 in 100, then the random chance for matching both becomes 1/100 * 1/100 = 1 in 10,000. The actual calculations are, of course, more sophisticated. Further, because these values become very large, AutoMatch reports them using Log Base 2. (A weight of 30 actually means 2 to the power of 30.)

Table 4 shows the distribution for ISCR-IHDD match weights in categories of PoO ranging up to 9,999,999-to-1 odds of separation between matches and non-matches.

Table 4. Posterior odds and distribution of matches

Posterior Odds	Posterior Odds Divided by Prior Odds	Convert to Log Base 2 to Obtain Weight	Matches in Weight Range*	Percent of Matches in Weight Range*
9,999,999 to 1	1,390,507,953,396	40.3	16,282	36.4%
999,999 to 1	139,050,670,194	37.0	10,124	22.7%
99,999 to 1	13,904,941,874	33.7	7,741	17.3%
9,999 to 1	1,390,369,042	30.4	4,620	10.3%
999 to 1	138,911,758	27.0	3,783	8.5%
99 to 1	13,766,030	23.7	2,146	4.8%

*Note: The ranges begin at the noted weight and extend up to the weight above. For example, in the bottom row (PoO 99-to-1) 2,146 matches were found with at least this weight of 23.7 and up to the next row weight of 27.0.

In this table, a large proportion of the matches are associated with high weight values, indicating a strong separation of matches from non-matches using the variables available. Of the total 44,696 matches, 36.4 percent have a PoO of at least 9,999,999 to 1. The median PoO for the entire distribution was 3,211,348 to 1, and the mean was 1,739,923 to 1.

Because the weights were so high, a cutoff value for all match passes of 23.7, or 99 to 1, was chosen. As shown in Table 4, this means that even the weakest category of matches had a PoO between 99 to 1 and 999 to 1, and this category itself was a small fraction of the total (2,146 matches or 4.8%).

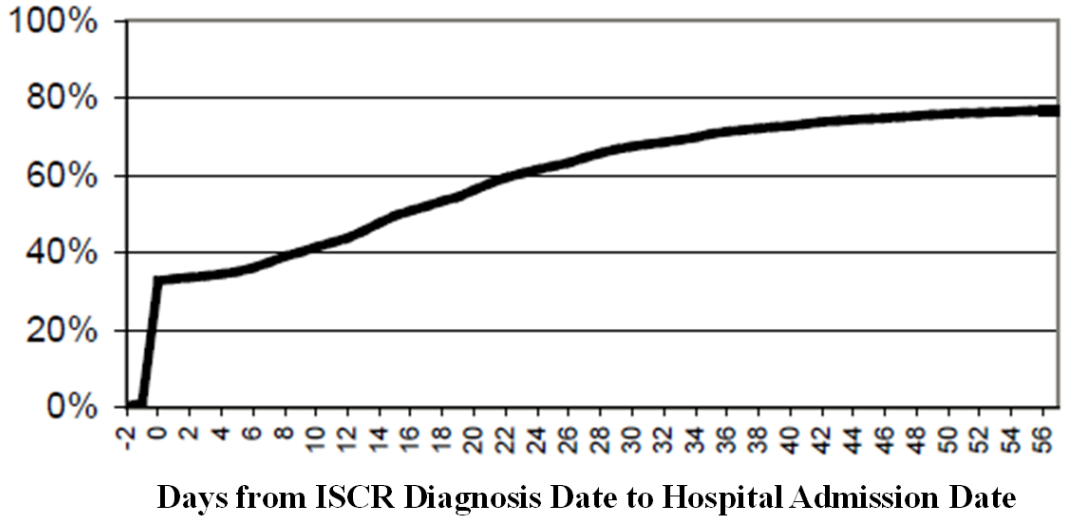
Evaluation of Linkage Results: After applying the software using the algorithm described above, a total of 29,381 ISCR records matched in a one-to-many relationship to 44,696 hospital records. The linkage can be evaluated from two perspectives.

Overall, these results are consistent with findings when using "weak" identifiers rather than strong identifiers such as person name or social security number. Such weak-identifier linkages tend to have lower match rates (lower sensitivity, e.g. 80%) but are more likely to find "true" matches (high specificity).[1,3]

A. ISCR Records that Matched

The overall match rate for ISCR records restricted to hospital admission dates up to 2 days prior to diagnosis date and 60 days after it was 76.8 percent (29,381 of 38,247 records). The cumulative distribution by the difference in days for the date from diagnosis to admission date was particularly interesting and is shown in Figure 5.

Figure 5. Cumulative percentage of matching ISCR records by days from ISCR diagnosis Date to hospital admission date



A large number of records (12,516) matched exactly on admission and diagnosis dates. Also the curve appears to flatten as the difference approaches 60 days.

B. Hospital Records That Matched

1. Table 5 shows specific match rates by type of diagnosis (primary, secondary, or history only) and procedures. As noted previously, hospital records included cases with only a history of breast cancer and these were less likely to match. The most likely records to match have either a principal or a secondary diagnosis of breast cancer, as well as a related procedure (biopsy, lumpectomy, or mastectomy). For these cases, the match rate was 75.7 percent (38,554 of 50,909 records). An additional 6,142 matches were found with no such procedures or history-only diagnosis, for a total of 44,696 matches.

Table 5. Match rates for hospital records

Procedures	History Only	Primary Diagnosis	Secondary Diagnosis	Primary or Secondary Diagnosis
None	0.9%	31.5%	7.9%	17.0%
Biopsy Only	7.8%	84.2%	61.2%	82.4%
Biopsy + Lumpectomy	4.5%	83.1%	57.7%	82.7%
Biopsy + Mastectomy	0 of 3 records	78.9%	35.9%	76.4%
Biopsy + Lumpectomy + Mastectomy	0 of 1 records	81.6%	0 of 0 records	81.6%
Lumpectomy only	5.3%	78.1%	49.4%	77.3%
Lumpectomy + Mastectomy	0 of 6 records	74.3%	64.0%	73.5%
Mastectomy Only	4.0%	65.0%	58.8%	64.9%
Any Above Procedure	5.6%	76.4%	55.5%	**75.7%**

2. Note Regarding Length of Stay. As a final note, it was interesting to consider this match relative to one performed nearly 10 years ago at IDPH that attempted to link ISCR to inpatient records only. At that time, the match rates were poor for inpatient length of stay (LOS) less than three days, but reasonable for longer LOS. Figure 6 shows the distribution of matched records for the current project by LOS as well as by inpatient or outpatient surgery.

Figure 6. Distribution of matched ISCR records by inpatient and outpatient surgery length of stay in days

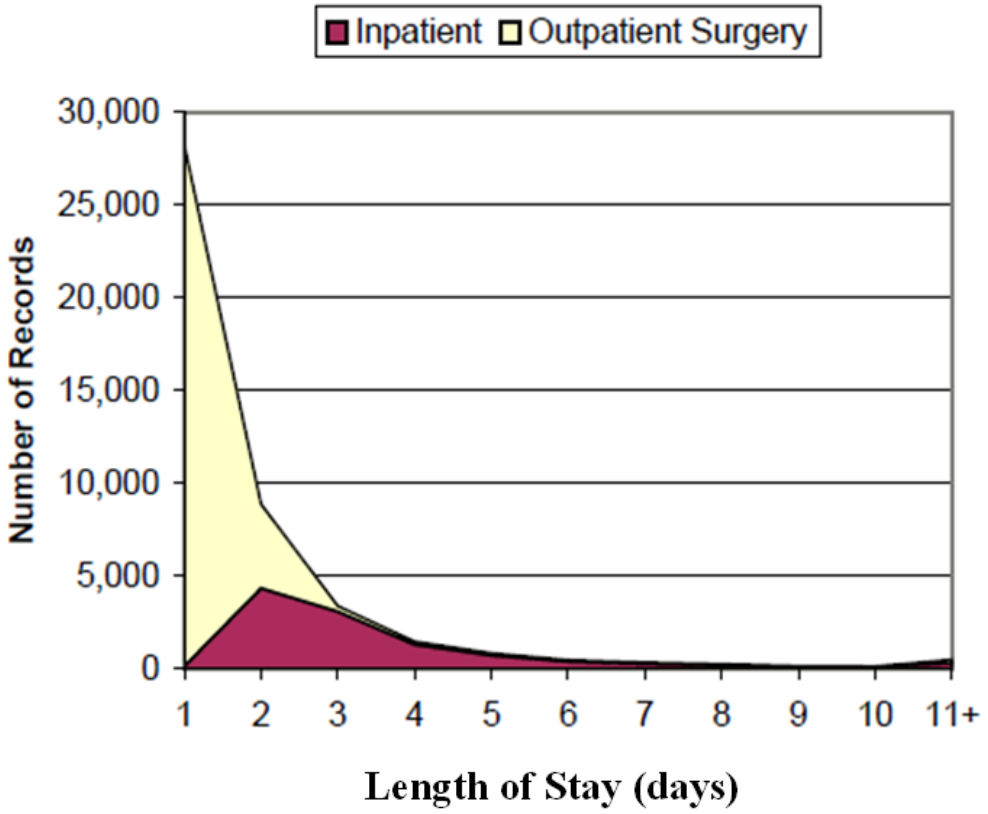


Figure 6 emphasizes the importance of outpatient surgery records in understanding the relationship between the cancer registry and hospital data.

References:

1. Cook LJ, Olson LM, and Dean JM. (2001). Probabilistic record linkage: relationships between file sizes, identifiers, and match weights. *Methods of Information in Medicine*, 40, 196-203.

2. Leonard DR, Rariden RL, Beccue B, and Shen T. (2006). Attempts to minimize manual review during registry deduplication. *Journal of Registry Management*, 33(1), 17-23.

3. Camelot Consulting. (2004). Is The Link King right for you? The Link King. Available at: http://www.the-link-king.com/LK_right.html. Accessed August 2, 2009.

Contributors to Appendix A

Missouri Cancer Registry and Research Center (MCR-ARC) Data Linkage Experience

Saba Yemane, BA, BS
Database Administrator
Missouri Cancer Registry and Research Center
University of Missouri School of Medicine
Department of Health Management and Informatics
P.O. Box 718
Columbia MO 65211
Phone: 573-884-6130
E-mail: yemanes@health.missouri.edu

Pennsylvania Cancer Registry Experience with Link Plus

Michelle L. Esterly, RHIA, CTR
Pennsylvania Cancer Registry Database Manager
Department of Health| Bureau of Health Statistics and Research
555 Walnut Street, 6th Floor
Harrisburg, PA 17101
Phone: 717-547-3697 | Fax: 866-531-8238
www.health.state.pa.us | PCR Website

The Illinois State Cancer Registry Data Linkage Experience

Lori Koch, BA, CCRP, CTR
Registry Manager
Illinois State Cancer Registry
Illinois Dept. of Public Health
535 W. Jefferson Street, 3rd Floor, Springfield, IL 62761
Phone: 217-785-7132 (Office line**)**
Phone: 217-557-4090 (Research line)
Fax: 217-557-5152

Hospital Discharge Linkage Methodology and Results From Massachusetts Cancer Registry (MCR)

Richard Knowlton, MS
Epidemiologist
Massachusetts Cancer Registry
E-mail: richard.knowlton@state.ma.us

Nebraska Cancer Registry Hospital Discharge Data linkage:

Ge Lin, Ph.D
GIS Program coordinator, Division of Public Health Nebraska Department of Health and Human Services
301 Centennial Mall S
Lincoln, NE 68509
Phone: 402-471-0920
Cell Phone: 402-416-8539
E-mail: ge.lin@nebraska.gov

Usefulness of Billing Data from Community Specialty Practices for Supplementing Treatment Reporting to Central Registries

Lynne Penberthy MD, MPH
Phone: 240-276-6864
E-mail: lynnepenberthy.schumacher-penberthy@nih.gov

Probabilistic Linkage Methodology for Match of Breast Cancer Cases July 2009

Therese A. Dolecek, Ph.D. and Daniel R. Leonard, M.S.
Contact: Therese A. Dolecek, Ph.D.
Research Associate Professor of Epidemiology Fellow Institute for Health Research and Policy School of Public Health MC 923 University of Illinois at Chicago
1603 W. Taylor St. Rm 883
Chicago, Illinois 60608
Phone: 312-996-9516
Fax: 312-996-0064
E-mail: tdolecek@uic.edu

Appendix B. Diagnosis Codes Used to Identify Breast Cancer

The ICD-9-CM diagnosis codes included in the HCUP Clinical Classification Software (CCS) category 24 (Cancer of breast) for females are available online at http://www.hcup-us.ahrq.gov/reports/statbriefs/sb15.jsp (accessed 7/6/09). They consist of the following:

– 174.0 Malignant neoplasm of female breast (nipple and areola)
– 174.1 Malignant neoplasm of female breast (central portion)
– 174.2 Malignant neoplasm of female breast (upper-inner quadrant)
– 174.3 Malignant neoplasm of female breast (lower-inner quadrant)
– 174.4 Malignant neoplasm of female breast (upper-outer quadrant)
– 174.5 Malignant neoplasm of female breast (lower-outer quadrant)
– 174.6 Malignant neoplasm of female breast (axillary tail)
– 174.8 Malignant neoplasm of female breast (other specified sites of female breast)
– 174.9 Malignant neoplasm of female breast (breast, female, unspecified)
– 233.0 Carcinoma in situ of breast
– V10.3 Personal history of malignant neoplasm, breast

Appendix C. Procedure Codes Associated with Breast Cancer

The ICD-9-CM procedural codes included in the HCUP Clinical Classification Software (CCS) categories associated with breast cancer are available online at http://www.hcup-us.ahrq.gov/reports/statbriefs/sb15.jsp (accessed 7/6/09). They consist of the following:

CCS procedure category 166 (Lumpectomy, quadrantectomy of breast)
– 85.20 Excision or destruction of breast tissue, not otherwise specified
– 85.21 Local excision of lesion of breast
– 85.22 Resection of quadrant of breast
– 85.23 Excision of nipple

CCS procedure category 167 (Mastectomy)
– 85.41 Unilateral simple mastectomy
– 85.42 Bilateral simple mastectomy
– 85.43 Unilateral extended simple mastectomy
– 85.44 Bilateral extended simple mastectomy
– 85.45 Unilateral radical mastectomy
– 85.46 Bilateral radical mastectomy
– 85.47 Unilateral extended radical mastectomy
– 85.48 Bilateral extended radical mastectomy

In addition, ICD-9 codes for biopsy were obtained from ICDData.com, a free web site listing available at http://www.icd9data.com/2007/. These included
– 85.11 Closed [percutaneous] [needle] biopsy of breast
– 85.12 Open biopsy of breast
– 85.19 Other diagnostic procedures on breast

For CPT procedural codes, HCUP makes downloads available for its CCS categories online at http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp (accessed 7/6/09). They consist of the following:

CPT Codes Start	End	CCS#	CCS Label
19120	19126	166	Lumpectomy, quadrantectomy of breast
19160	19162	166	Lumpectomy, quadrantectomy of breast
19297	19297	166	Lumpectomy, quadrantectomy of breast
19301	19302	166	Lumpectomy, quadrantectomy of breast
19180	19240	167	Mastectomy
19300	19300	167	Mastectomy
19303	19307	167	Mastectomy

CPT codes for biopsy were obtained from two sources
1) Illinois's Breast and Cervical Cancer project, available online at http://www.lchd.us/PDF/2008%20IBCCP%20CPT%20Codes.pdf (accessed 11/24/2014).

2) Ohio's Breast and Cervical Cancer project, available online at http://www.healthy.ohio.gov/~/media/HealthyOhio/ASSETS/Files/bccp/2014%20Ohio%20BCCP%20CPT%20Codes%20Feb%2014.ashx (accessed 11/25/2014).

Breast biopsy codes from both of these sources included: 19100,19101,19102, and19103. The 2014 version of the Ohio study deleted codes 19102 and 19103.

Appendix D. Linkage Algorithm Used in Automatch

```
PROGRAM GEOMATCH MULTIPLE
DICTA hosp
DICTB iscr
;
;
BLOCK1 CHAR hosp hosp
BLOCK1 CHAR ZIP ZIP
BLOCK1 CHAR DOB DOB
;
MATCH1 CHAR hosp hosp 0.89 0.01
MATCH1 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH1 DATE8 admdate datedx 0.99 0.1 3 90
MATCH1 CNT_DIFF zip zip 0.99 0.01 1
;
CUTOFF1 23.9 23.7 23.9
;
;
BLOCK2 CHAR hosp hosp
;
MATCH2 CHAR hosp hosp 0.89 0.01
MATCH2 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH2 DATE8 admdate datedx 0.99 0.1 3 90
MATCH2 CNT_DIFF zip zip 0.99 0.01 1
;
CUTOFF2 23.9 23.7 23.9
;
;
BLOCK3 CHAR DOB DOB
;
MATCH3 CHAR hosp hosp 0.89 0.01
MATCH3 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH3 DATE8 admdate datedx 0.99 0.1 3 90
MATCH3 CNT_DIFF zip zip 0.99 0.01 1
;
CUTOFF3 23.9 23.7 23.9
;
;
BLOCK4 CHAR zip zip
;
MATCH4 CHAR hosp hosp 0.89 0.01
MATCH4 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH4 DATE8 admdate datedx 0.99 0.1 3 90
MATCH4 CNT_DIFF zip zip 0.99 0.01 1
;
```

```
CUTOFF4 23.9 23.7 23.9
;
;
BLOCK5 CHAR admdate datedx
;
MATCH5 CHAR hosp hosp 0.89 0.01
MATCH5 CNT_DIFF DOB DOB 0.99 0.01 1
MATCH5 DATE8 admdate datedx 0.99 0.1 3 90
MATCH5 CNT_DIFF zip zip 0.99 0.01 1
;
CUTOFF5 23.9 23.7 23.9
;
vartype admdate critical
vartype dob critical
vartype zip critical
```