Building a Virtual Cancer Research Organization

Mark C. Hornbrook, Gene Hart, Jennifer L. Ellis, Donald J. Bachman, Gary Ansell, Sarah M. Greene, Edward H. Wagner, Roy Pardee, Mark M. Schmidt, Ann Geiger, Amy L. Butani, Terry Field, Hassan Fouayzi, Irina Miroshnik, Liyan Liu, Robert Diseker, Karen Wells, Rick Krajenta, Lois Lamerato, Christine Neslund Dudas

Background: **The Cancer Research Network (CRN) comprises the National Cancer Institute and 11 nonprofit research centers affiliated with integrated health care delivery systems. The CRN, a public/private partnership, fosters multisite collaborative research on cancer prevention, screening, treatment, survival, and palliation in diverse populations.** ***Methods:*** **The CRN's success hinges on producing innovative cancer research that likely would not have been developed by scientists working individually, and then translating those findings into clinical practice within multiple population laboratories. The CRN is a collaborative virtual research organization characterized by user-defined sharing among scientists and health care providers of data files as well as direct access to researchers, computers, software, data, research participants, and other resources. The CRN's research management Web site fosters a high-functioning virtual scientific community by publishing standardized data definitions, file specifications, and computer programs to support merging and analyzing data from multiple health care systems.** ***Results:*** **Seven major types of standardized data files developed to date include demographics, health plan eligibility, tumor registry, inpatient and ambulatory utilization, medication dispensing, laboratory tests, and imaging procedures; more will follow. Data standardization avoids rework, increases multisite data integrity, increases data security, generates shorter times from initial proposal concept to submission, and stimulates more frequent collaborations among scientists across multiple institutions.** ***Conclusions:*** **The CRN research management Web site and associated standardized data files and procedures represent a quasi-public resource, and the CRN stands ready to collaborate with researchers from outside institutions in developing and conducting innovative public domain research. [J Natl Cancer Inst Monogr 2005;35:12–25]**

Multi-Institutional Cancer Research

Cancer researchers have a long tradition of multisite collaborative clinical trials, epidemiologic studies, and other research studies. Multisite studies allow accessing large populations to provide sufficient numbers of eligible research participants and to exploit existing secondary data. Most cancer studies are unique in aims, hypotheses, and data sets, increasing the barriers to pooling data across studies. In response to these barriers, the National Cancer Institute (NCI) has invested heavily in research programs that standardize data collected from multiple sites, including the National Institutes of Health (NIH) Roadmap initiative (http://nihroadmap.nih.gov/).

To pursue consistent national estimates of changes over time in the incidence and economic burden of cancer in the United States, the NCI created the Surveillance of Epidemiology and End Results (SEER) system to define and collect standardized information on tumors in representative populations *(1)*. NCI has enhanced the SEER system by linking Medicare claims data to tumor registry data for those Medicare beneficiaries who appear in the SEER system *(2)*. Because Medicare facility and professional claims are already in a nationally standardized format, the data are ready to use once the linkage is made using Social Security numbers (Medicare Health Insurance Claims [HIC] numbers). Large numbers of publications have been produced using the SEER and SEER-Medicare databases (see http://seer.cancer.gov/).

The SEER–Medicare linkage provides a useful picture of time-series and cross-section patterns of cancer rates and cancer-related patterns of care for Medicare beneficiaries. Nevertheless, it does not cover Medicare Working Aged beneficiaries, working-age adults, children, or older persons who never established Medicare eligibility. To obtain access to diagnosis, utilization, and expense data on non-Medicare persons with cancer, NCI established the Cancer Research Network (CRN) in 1998. One goal of the CRN is to develop a decentralized data standardization process to support pooling of clinical, utilization, and administrative data across multiple integrated delivery systems over time.

The CRN comprises the National Cancer Institute and 11 nonprofit research centers affiliated with integrated health care delivery systems. The CRN fosters multisite collaborative research on cancer prevention, screening, treatment, survival, and palliation in diverse populations. This overarching aim of the CRN fosters efficient and effective research on variations in cancer prevention and treatment policies and practices *(3–5)*.

To date, the CRN has launched 27 research projects. These projects have employed a variety of data collection modalities,

Affiliations of authors: Center for Health Research, Northwest/Hawaii, Kaiser Permanente Northwest, Portland, OR (MCH, DJB, GA); Center for Health Studies, Group Health Cooperative, Seattle, WA (GH, SMG, EHW, RP); Clinical Research Unit, Kaiser Permanente Colorado, Denver, CO (JLE); Center for Health Research, Kaiser Permanente Hawaii, Honolulu, HI (MMS); Office of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA (AG); HealthPartners Research Foundation, Minneapolis, MN (ALB); Meyers Primary Care Trust/Fallon Health Systems, Worcester, MA (TF, HF); Harvard Pilgrim Health Care, Boston, MA (IM); Division of Research, Kaiser Permanente Northern California, Oakland, CA (L. Liu); Department of Prevention and Research, Kaiser Permanente Georgia, Atlanta, GA (RD); Center for Health Services Research, Henry Ford Health System, Detroit, MI (KW, RK, L. Lamerato, CND).

Correspondence to: Mark C. Hornbrook, PhD, Center for Health Research, 3800 N. Interstate Ave., Portland, OR 97227 (e-mail: mark.c.hornbrook@kpch.org).

See "Notes" following "References."

DOI: 10.1093/jncimonographs/lgi033

including qualitative and quantitative surveys of enrollees, providers, and health plan leaders; reviews of enrollee medical records; laboratory testing of tissue specimens; and collection and aggregation of automated (electronic) data. As of this writing, 25 papers have been published from CRN projects. Examples of CRN studies include analyses of tobacco control policies and interventions in integrated delivery systems *(6–9)*, rates of failure of breast cancer screening procedures *(10)*, breast and cervical cancer screening guidelines and processes *(11–14)*, patterns of use of hormone replacement therapy over time *(15)*, variations in HMO disenrollment among cancer patients by race/ethnicity *(16)*, efficacy of bilateral prophylactic mastectomy among women with elevated breast cancer risk *(17)*, and research methods *(18–23)*.

The CRN's success hinges on producing innovative cancer research that would not likely have been developed by scientists working independently and translating research findings into clinical practice within multiple population laboratories. The CRN's "virtual research organization" *(24)* promotes flexible, secure, coordinated resource sharing. Resource sharing, defined by the user community, includes data file exchange as well as direct access to collaborators, computers, software, data, research participants, and other resources required to develop, conduct, and disseminate the results of research on cancer. The CRN's research management Web site fosters a high-functioning virtual scientific community.

In this article, we highlight development by the CRN of standardized data definitions, file specifications, and computer programs to support combining data from multiple health care systems. Types of data to date include tumor registry, inpatient and ambulatory utilization (claims and encounters), diagnoses, procedures, dispensing, laboratory tests, and imaging procedures; more will follow. We also discuss the regulations and procedures governing access to health plan data for research. Finally, we describe how the CRN envisions fitting in with the NIH Roadmap initiatives and the developing national standards for research and care, and in particular how the CRN data standardization process will align with the cancer Biomedical Informatics Grid (caBIG) efforts.

Methods

Managing Health Plans' Burden of Information Disclosure

Health plans have both proprietary interests and regulatory compliance goals with respect to disclosing their data for public domain research. When viewing a request to participate in public domain research, health plan executives balance the potential benefits to the health plan, patients, and society against the potential risks to the plan and its members from disclosure of proprietary information and unauthorized disclosure of protected health information. In the case of the CRN health plans, their nonprofit status carries the added responsibility of generating societal benefit. Public domain research is one avenue for discharging this responsibility. The CRN's design systematizes and standardizes collaborative research activities across the participating health plans.

The CRN provides three proprietary benefits to participating health plans: ability to pool data from multiple health plans so that the identities and attributes of any single health plan are masked in research publications (institutional confidentiality), a process for screening research proposals to ensure that they represent legitimate scientific research, and an NIH Certificate of Confidentiality that protects research data from disclosure (patient confidentiality).

HIPAA and Human Subjects Compliance

CRN researchers have an inherent advantage as health plan or medical group employees, in that they can access individual patient-level data and link data from multiple legacy data systems for research purposes. Primary data collection activities for research projects that are covered by carefully constructed, signed informed consent forms permit creation of person-level databases that can be placed in the public domain. By contrast, research databases created with public funding without the specific written consent of persons whose data are contained therein must meet the criterion of minimal risk and potential social benefit.

Health plans also must comply with the regulations issued to implement the Health Insurance Portability and Accountability Act of 1996 (PL 104–191) (HIPAA). HIPAA defines protected health information (or PHI) as identifiable data on an individual's health status, medical care, medical history, and behaviors. In general, HIPAA requires express patient authorization for the use or disclosure of PHI in research activities.

Health plans are defined as "covered entities" under HIPAA, and as such, each organization is responsible for preventing unauthorized disclosure and use of protected health information. With appropriate permissions granted by research participants, individuals' data can be linked to external databases, such as birth and death certificates and Medicare and Medicaid claims. For research projects judged by institutional review boards (IRBs) to provide scientific benefits with only minimal risk, the consent form requirement can be waived, thereby avoiding selection bias associated with the decision to volunteer for research.

Best practices for human subjects' protection are shared among the CRN sites. All CRN sites operate their own IRBs and require every proposal that involves data on their members, employees, or providers to be reviewed locally. The CRN provides a clearinghouse for IRB application forms, schedules, and procedures.

Transferring person-specific research data from a CRN site to another research site requires approval of IRBs, HIPAA privacy officers, HIPAA data security officers, and research center directors. The CRN established a secure data transfer Web site that meets HIPAA and Medicare data security standards so that every keystroke, starting with the login sequence, and all attached data files are encrypted. Moreover, the site requires users to install security certificates and to have user accounts and passwords.

To preserve patient confidentiality, the CRN relies on tables of anonymous data to distribute information to external audiences. Health plan data are aggregated to a level at which it is virtually impossible to reconstruct observations on any individual contained in the analysis data set. The CRN Cancer Counter (see further description later) is an example of enabling access to cancer incidence data by imposing a data query engine between the user and the raw data so that frequency counts fewer than five can be suppressed.

We define "virtual data" as the decentralized standardized files held by individual sites, ready to process when a standard data query program is submitted for a specific purpose. The CRN developed descriptions of automated data systems, standardized file specifications, and standard computer programs to be run

against standardized files to ensure that users would not create increased threats to data security inside health plan IT firewalls. Health plan legacy files and variables are described generically, not literally, in CRN documentation. Standardized data files can be combined across multiple sites to produce larger sample sizes with more demographic and health system diversity than is possible from a single site.

Scientific and Data Resources Core (SDRC)

Every health plan has its own variations on internally developed and vendor-supported data systems. The newest wave of IT development is electronic medical records (EMRs). All the HMOS of the CRN have or are implementing EMRs. Ten of the 11 HMOs are using the EPICCare EMR product.

The business imperatives of integrated health care delivery systems mean that the data systems of any large health plan contain comparable information, such as diagnoses, procedures, encounters, and dispensings. Similarities in business operations provide the opportunity to extract common variables across health plans.

The CRN has assembled information about each CRN health plan and its computerized data systems. An obstacle to this work is the lack of standardization among data system vendors, operating systems, software, file layouts, and file content. The CRN SDRC was created in part to build on previous work that described the data systems of the participating health plans *(25–27)*. One of its charges was to increase the quality and efficiency of CRN research projects by identifying and disseminating optimal methods for data collection, management, and transfer. The SDRC is comprised of site data managers (SDMs) and data principal investigators (PIs) who have comprehensive knowledge of the data-related capabilities of their sites.

Heterogeneous Structures of HMO Automated Data Systems

The health plans participating in the CRN are heterogeneous on many dimensions, including ownership, size, structure, information systems, and years in operation. Of crucial importance to data standardization efforts are the interrelationships among age of health plan and size and structure of local information systems. One crucial distinction is whether the HMO has encounter data systems, claims data systems, or both as the primary structure of its data warehouses. The encounter versus claims mapping issue is relevant both within and across health plan data systems.

Encounter data systems often evolve not from a need to bill for services but, rather, from a need to document utilization for operations management purposes. Because there is no standardized format for encounter systems, it may be difficult to obtain comparable data quality across sites and over time.

Claims data systems have three streams—facility bills in UB-92 format, professional bills in HCFA-1500 format, and drug claims in a standardized pharmacy benefits management format. Claims data systems evolve from a need to receive, adjudicate, and pay bills for health care services.

Documenting Legacy Data Systems

Each local SDM must have extensive knowledge of his or her site's data systems and know whom to contact to obtain pertinent information. The SDMs also should be familiar with all the lines of business for their respective health plans, as well as the structure of the health care delivery system. For instance, some health plans have internal home health agencies, and all of the home visits are documented in a computerized home health record system. In contrast, other health plans use community home health agencies exclusively and may have little information on what services are delivered by these agencies, particularly for hospice patients. Some plans have claims data from home health agencies, and others have no data at all.

This level of detailed understanding of the scope of services produced internally by each health plan and the overlapping or complementary use of outside contractors for specified types of services is essential to interpreting clinical, utilization, and expense data for each health plan and across health plans. A programmer who is not familiar with a health care organization often cannot detect implausible or misleading patterns in its data. A health care analyst with limited programming skills requires considerable technical assistance to define and create standardized data files for his or her health plan. Hence, either SDMs need to have both programming skills and health care organizational knowledge or assistance is needed to effectively gather data from their system and prepare data for pooling across multiple plans.

For each element within each content area, idiosyncrasies specific to that site should be documented. For example, at some sites, membership data from before a certain date may be unreliable. At another site, automated medical record systems may have been implemented that changed the quality or quantity of diagnostic coding before the implementation of the EMR. Another site might use a mixture of standard and "homegrown" procedure codes. Part of the process of creating a standardized data structure is building and continually updating the documentation of each data system in the participating health plans.

Establishing Standardized Data Files

The CRN leadership conceived the approach of using Web publication of standardized data file specifications and building local versions of standardized files as mechanisms to produce comparable data across sites for purposes of proposing or conducting research. Original legacy files and local versions of standardized files remain at the local sites. Standardized extraction files are distributed from a central source to be run locally against standardized files, and then the output files are transferred via secure data transfer to the requesting site. A schematic representation of the process for building local standardized files is shown in Figure 1.

Content areas and data elements that are commonly required for research studies are identified, and standardized data dictionaries are created for each of the content areas, specifying a common format for each of the elements—variable name, variable label, extended definition, code values, and value labels. This allows SDMs to construct comparable data sets using potentially different sources and formats. Our vision is that using standardized files to manage the interfaces between project data needs and health plan legacy information systems, and between project programmers and health plan programmers, will become a "best practice" for all CRN research projects.

Data standardization involves the following steps: specifying common variable names, labels, coding, and definitions; writing programs to extract variables stored in HMO legacy information

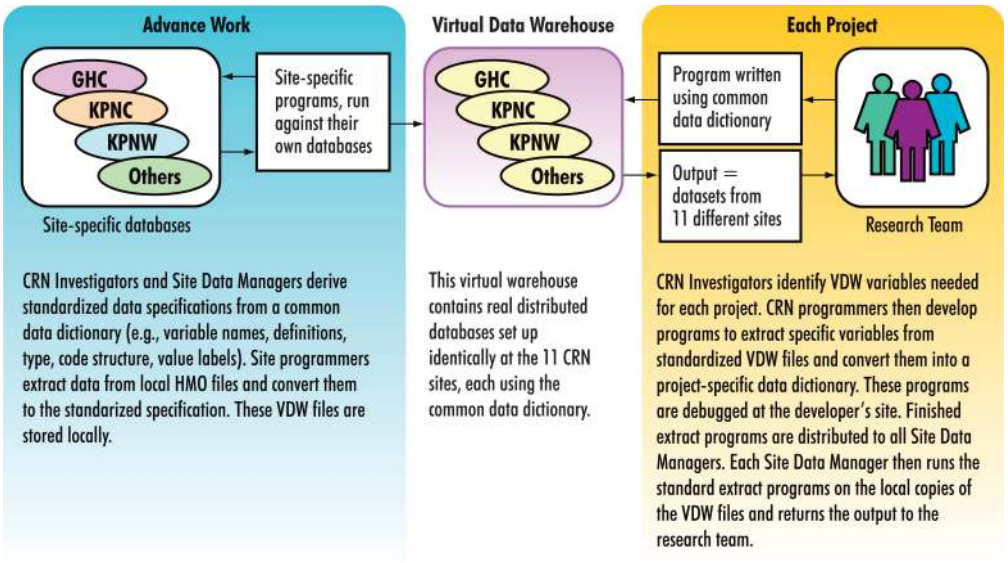

Fig. 1. Schematic of the CRN standardized data warehouse.

systems and convert them to the common standards; testing standardized data for consistency and accuracy; and teaching researchers and their programmers how to use the standardized files to guide construction of analysis files for approved research projects. The CRN data standards are derived from the information system standards contained in the standards for accreditation of health care networks (integrated delivery systems, managed care organizations, preferred provider organizations, etc.) by the Joint Commission on Accreditation of Healthcare Organizations (http://www.jcaho.org/accredited+organizations/health+care+network/network+accreditation.htm). Across the CRN sites, we find widespread use of ICD-9-CM diagnosis and procedure coding systems, CPT-4 procedure coding systems, HCPCS procedure coding systems, and ICD-O3 oncology coding systems. Some health systems use SNOMED CT (supported by the College of American Pathologists), which encompasses all of the above systems. The pharmaceutical coding system is the National Drug Code, supported by the U.S. Food and Drug Administration. Systems for standardizing clinical terminologies across nations, clinical specialties, health care facilities, and health care systems include Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/), Logical Identifiers Names and Codes (LOINC, http://www.nlm.nih.gov/research/umls/loinc_main.html), and HL7 (http://www.hl7.org/). The CRN will be affected by the evolution of these common terminology and coding systems as our host health plans adopt them in their EMR systems and other automated information systems.

The CRN Project Leaders Forum designates content areas that it believes would be useful for cancer research, such as tumor registry, enrollment, and utilization. With input from site researchers, members of the SDRC with content experience or interest discuss which data elements are commonly required for research studies and are likely to be found across health plans. The number of elements included in each content area must be large enough to be useful for research but not so large that creating and using standardized files becomes too unwieldy. For example, enrollment information might include employer, benefit, and family relationship information and consist of dozens of fields. Standardized data elements selected for this content area were the more commonly needed fields of patient identifier, enrollment by month and year, and type of payer (e.g., Medicare). The data dictionary consists of variable names, definitions, and formats, as well as information relevant to the availability, reliability, and validity of each variable. The specifications for the standardized demographics file are shown in Table 1.

Other standardized file specifications have been developed for health plan enrollment and benefit information (see Table 2), tumor registry systems (see Table 3), dispensings (see Table 4), laboratory tests (see Table 5), and utilization (encounters, stays, diagnoses, procedures, and provider specialty; see Table 6). The availability of these data by year across the participating health plans is shown in Table 7.

Each SDM writes the necessary code to extract information from health plan data systems and convert it into a file that matches the standardized files as closely as possible. If any variable is not available from health plan systems, that information is added to the documentation for the standardized files. If a match is not straightforward, that information is also documented to alert potential future users of validity and consistency problems when using or pooling data from multiple plans.

Once the appropriate legacy systems have been identified for a given content area, the SDMs, using site-specific operational definitions, are responsible for writing and testing programs to extract the raw data and convert them into the format prescribed by the data dictionary. As the source systems are likely to be different at each site, these programs may be very different from site to site. They all, however, should yield comparable data. Depending on the content area and the SDM's familiarity with the content area, this phase might be fairly time consuming. Quality or reasonableness checks should be done at each step to ensure that the content of the standardized files is what is intended. Each site maintains a central repository of the finalized programs, along with documentation of special issues.

Setting up the Local Standardized Data Sets

The CRN has two models for implementing data standardization, allowing for flexibility in working under the resource constraints faced by each site. Under one model, the programs that extract standardized data are tested and debugged, but no

Table 1. CRN standardized demographics data structure*

Variable Name	Variable Definition	Values	Comments
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
BIRTH_DATE	Numeric 4	SAS Date	
GENDER	Char 1	M = male F = female O = Other T = Transsexual U = Unknown	Using SEER gender categories
RACE1 RACE2 RACE3 RACE4 RACE5	Char 2	'01' = "White" '02' = "Black" '03' = "American Indian, Aleutian, or Eskimo" '04' = "Chinese" '05' = "Japanese" '06' = "Filipino" '07' = "Hawaiian" '08' = "Korean" '09' = "Asian Indian, Pakistani" '10' = "Vietnamese" '11' = "Laotian" '12' = "Hmong" '13' = "Kampuchean" '14' = "Thai" '20' = "Micronesian, NOS" '21' = "Chamorran" '22' = "Guamanian, NOS" '25' = "Polynesian, NOS" '26' = "Tahitian" '27' = "Samoan" '28' = "Tongan" '30' = "Melanesian, NOS" '31' = "Fiji Islander" '32' = "New Guinean" '96' = "Other Asian, incl. Asian, NOS and Oriental, NOS" '97' = "Pacific Islander, NOS" '98' = "Other" '99' = "Unknown"	
HISPANIC	Char 1	'Y' = Yes 'N' = No ' ' = Unknown	Hispanic origin (ethnicity)

*Code Race according to the SEER Race categories and definitions, as defined at: http://www.seer.cancer.gov/tools/codingmanuals/race_code_pages.pdf. If a person's race is recorded as white and any other race, code to the appropriate other race first then code to white in the next field. If a person's race is recorded as a combination of Hawaiian and any other races, code Race1 as Hawaiian then code race 2–5. Otherwise, code Race1 to the first stated non-white race ('02'–'98'). If a specific Asian code is available, do not use '96' Asian NOS. If Race1 is unknown '99', then Race 2–5 must also be '99'. If only 1 race is reported for the person use code '88' for the remaining fields.

standardized data files are created until a specific data request is submitted to be run against the standardized files. This model may be used in sites for which the costs of extracting and storing data are relatively high.

The other model involves extracting and translating all of the data elements for all members in advance into local standardized files. This model requires more resources up front, but it substantially improves turnaround time. In addition, it ensures that these data elements will be preserved, regardless of what may happen to any site's legacy systems over time.

The CRN is developing site-specific 200 000-person samples for each of the standardized data files. Uniform sample sizes

Table 2. CRN standardized enrollment data structure

Variable name	Variable Definition	Values	Comments
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
ENR_MONTH	Numeric 4	1–12	
ENR_YEAR	Numeric 4	Values of the form 1980	Whatever time period works at the HMO
INS_MEDICARE	Insurance Medicare Char 1	"Y"=Yes. " " = No or missing	
INS_MEDICAID	Insurance Medicaid Char 1	"Y"=Yes. " " = No or missing	
INS_COMMERCIAL	Insurance Commercial Char 1	"Y"=Yes. " " = No or missing	
INS_PRIVATEPAY	Insurance Private Pay Char 1	"Y"=Yes. " " = No or missing	
INS_OTHER	Insurance Other Char 1	"Y"=Yes. " " = No or missing	

Table 3. CRN standardized tumor registry data structure

Variable Name	Description	Format	Comments
ICD-O3 site	Cancer site (e.g. breast, prostate)	Char(4) Example: C619	From ICO-O Version 1—1976–1989 Version 2—1990–2000 Version 3—2001
StageGen	General Stage: 0 = In situ 1 = Localized 2 = Regional by direct extension 3 = Regional to lymph nodes 4 = Regional both direct extension and lymph nodes 5 = Regional, NOS 7 = Distant metastasis 9 = Unstageable, unknown, unspecified B = Benign	Char(1)	Some sites may have a different coding scheme for these—need to assure consistency at the lowest common denominator
StageAJ	AJCC summary stage or "best AJCC stage" 0, 0a, 0is (URINARY TRACT SITES) 1, 1A, 1B, 1C, 1S (TESTIS), 1A1, 1A2, 1B1, 1B2 2, 2A, 2B, 2C, 3, 3A, 3B, 3C 4, 4A, 4B, 4C	Char(3)	Site-specific schemes apply using the American Joint Commission on Cancer Staging Manual. Versions vary over time (currently on v5). Best AJCC stage refers to the best of clinical or pathological stage, which means that if only clinical stage is available, that is used, but if both clinical and pathological are available, pathological is best. If only pathological stage is available, this is best AJCC.
AJCC_Ed	AJCC Staging Scheme Edition 0 = Not staged 1 = First Edition 2 = Second Edition 3 = Third Edition 4 = Fourth Edition 5 = Fifth Edition 6 = Sixth Edition 8 = Not applicable (no AJCC scheme) 9 = Unknown edition	Char(1)	
Morph	Morphology/histology (tissue type of cancer)	Char(4)	ICO-O Version 1—1976–1989 Version 2—1990–2000 Version 3—2001
Behavior	Behavior 0 = Benign 1 = Uncertain behavior, low malignancy potential, uncertain malignancy potential 2 = In situ 3 = Malignant, primary site 6 = Metastatic site 9 = Unknown metastatic or primary site	Char(1)	Benign lesions may be included in tumor registry if a "Reportable by Agreement" list exists for that particular site requesting registration of benign tumors of interest to clinicians or researchers. Rarely or never used (but allowable) are 6 (metastatic site) and 9 (unknown whether primary or metastatic).
Grade	Histologic grading and differentiation 1 = Well differentiated 2 = Moderately differentiated 3 = Poorly differentiated 4 = Undifferentiated, anaplastic 5 = T-cell 6 = B-cell 7 = Null cell (non T or B cell) 8 = NK cell (natural killer cell) 9 = Grade or differentiation not determined, not stated or not applicable	Char(1)	From ICD-O This field can be used to denote cell lineage for leukemias and lymphomas. This designation is 5, 6, 7, and 8 and used only for leukemia and lymphoma.
DXDate	Diagnosis date	SASdate Integer(5)	SASdate is Julian date. If a valid date is not available (say, the month and year are available, but not the day), a date must be forced to get a SASdate.
DXYear	Year of diagnosis	Numeric(4)	
DXAge	Age at diagnosis	Numeric(3)	
BDate	Birth Date	SASdate integer(5)	SASdate is Julian date. If a valid date is not available (say, the month and year are available, but not the day), a date must be forced to get SASdate.

Table 4. CRN standardized outpatient pharmacy data structure

Variable name	Variable Definition	Values	Comments
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
RXDATE	Date of dispensing	Type : numeric (4)	SAS date variable
NDC	National Drug Code	Char (11)	Please expunge any placeholders (ex. '-' or extra digit)
RXSUP	Days supply	Num (4)	
RXAMT	Amount dispensed	Num (4)	Number of units (pills, tablets) dispensed. Net amount per day per NDC.
RXMD	Prescribing MD	Character. Unique to each HMO.	Optional field. Use same coding scheme as PROVIDER in Utilization table
		EVERNDC Data Sub-Structure*	
NDC	National Drug Code	Char (11)	Please expunge any placeholders (ex. '-' or extra digit)
GENERIC	Generic Name	Char(105)	
BRAND	Brand Name	Char(100)	
GPI	Generic Product index	Char(14)	Optional, but may be useful if you have it
AHFS Therapeutic Class Code	American Hospital Formulary Service	Char(6)	Optional, but may be useful if you have it

*EVERNDC is a look-up table to identify drug products and supplies that have been introduced to and withdrawn from the market over the course of a study period.

across health plans are useful when pooling data from multiple health plans, because users cannot derive the identities of health plans by counting relative numbers of cases from each site; moreover, each health plan is equally represented in the analyses.

After any type of standardized files is created, a series of data edit programs are sent to all the sites to run against their local standardized file. The output tables are sent to all the CRN SDMs for conducting quality assurance and data editing reviews. These programs look for out-of-range codes, missing data, implausible data patterns, and unusual data patterns across site and time. Our goals in this step are to reveal cases of local programmers applying varying interpretations of the standardized file specifications, inability of some sites to match the standard specifications, or differences in patterns of care or documentation in the legacy systems from which local programmers extracted their standardized data. Even with these data integrity protections, local incentives and practices may lead seemingly comparable data to vary in comparability. For example, we know that diagnosis and service capture rates vary across health plans. Even when users make sure they are capturing the same data, capture rates might make seemingly comparable data less comparable.

Analysis of Standardized Data

CRN projects have two strategies for using the standardized data files. One strategy is to write a series of data processing and statistical analysis programs that are distributed to each site and run against the local versions of the standardized files, with the results transmitted back to the requesting site. This strategy works well for computing new variables (such as an annual comorbidity index for each person), creating pooled analysis files across multiple sites, and conducting descriptive analyses that can be accomplished by pooling the same data summary table from each site.

The second strategy follows from the pooled data file strategy, in which all the data needed for a research project are transferred to the PI's site and analyzed at that site. With pooled data sets, additional data quality checks are needed to ensure that interrelationships among selected variables are interpretable within each site's data; for example, rates of hospital admissions by age group and gender. A number of cross-tabulations should be performed on site-specific data to see whether the patterns are consistent, and if they are different, explainable—errors are possible at any stage of the data collection, extraction, merging, and analysis process. Thus, even when the pooled analysis file is created, users should be prepared to return to an earlier step in the process to correct errors and repeat the rest of the process using the corrected data.

If the local standardized files are already created and debugged, and data edits are completed, the files become an efficient means of answering queries to support proposal writing (e.g., estimating eligible study populations) and extracting analysis files for research projects. It is likely that many projects will have additional data needs that go beyond the variables contained in the standardized files. Nevertheless, the standardized files reduce overall data processing costs and enable programmers to focus their efforts on extracting consistent measures of the new study-specific variables. The CRN data structures are planned to be flexible and allow some of these variables to be incorporated into the standardized files if they are found to be reliable, valid, and useful for research.

Data Query Tools

The CRN has developed the "Cancer Counter" to support queries about incidence of primary tumors among members of the participating health plans. The Cancer Counter supports query definition on the following dimensions: tumor site, behavior, morphology, stage, health plan, vital status, race, gender, and Hispanic ethnicity. Once a population is selected, the Cancer Counter allows users to select one-way and two-way frequencies of the above variables. The Cancer Counter incorporates HIPAA protection by replacing cell counts less than five with a message indicating such, reducing the chances of identifying any person by linking his or her data to outside public databases.

Sites can support the counter by submitting detailed frequency tables and cross-tabulations, which are loaded into data files supporting the query tool. Alternatively, they can submit an individual level data file on tumors and the CRN Web site programmers will load the data into the Cancer Counter. The latter method requires IRB and HIPAA approvals, which many sites are willing

Table 5. CRN standardized laboratory procedures data structure*

Variable Name	Description	Format	Comments	Include in Lab VDW (Y/N)
MRN	Unique patient identifier within a practice site	Char(XX)	Length should be consistent with that of other VDW (able to accommodate all sites)	Y
LABDATE	Date of lab test—generally refers to date specimen was collected	SASdate integer (5)		Y
TESTCODE	Lab test code—used to relate to lab test description in reference table	Char(6)	Links to lab test description in reference table	Y
ABN_IND	Abnormal indicator—flag that designates abnormal results Y = abnormal Null = not abnormal, no information	Char(1)	Not applicable to all lab tests	Y
BATT_CDP	Battery code primary—designates the applicable battery code for tests within a given battery	Char(6)	Not applicable to all lab tests Useful but may not be practical due to individual site coding variation (e.g. This may be a "Fordism")	N
BATT_CDS	Battery code secondary—designates the applicable secondary battery code for tests within a given battery	Char(6)	Not applicable to all lab tests	N
COMP_CD	Company—designates the lab processing the specimen	Char(4)	Accommodates lab specimens sent to outside labs	N
DEPT_CD	Department code—refers to type of laboratory in which test was performed: BB = Blood bank C = Chemistry CS = Special Chemistry SO = Send outs	Char(4)	Not sure this is useful for VDW, more of an administrative tool within lab sites	N
DEPT_NM	Department name—description for department code (above)	Char(12)		N
RES_NON	Non-numeric test results—text results for tests (e.g. positive, negative, <.01, non-reactive)	Char(200)	For some labs this is the only type of valid results or reported results	Y
NORM_RAN	Normal range—expected range of values for specific test for specific type of pt	Char(15)	Needed to determine how individual patient's result compares to normal for a similar kind of pt	Y
NORM_L	Normal range low—lowest numeric value of normal range for specific test for specific type of pt	Decimal(11,4)		Y
NORM_H	Normal range high—highest numeric value of normal range for specific test for specific type of pt	Decimal(11,4)		Y
NUMERIC	Numeric test result—actual value of lab test	Decimal(11,4)		Y
ORD_DR	Ordering doctor number—site-specific doctor billing code	Char(5)	Size of field could vary across sites	N
DOC_NM	Ordering doctor name—relates to ordering doctor number	Char(25)		N
DOC_SPEC	Ordering doctor specialty (see specialty/department look-up table)	Char(3)	Probably site-specific, but could be translated to a standard scheme across CRN sites	Y
DOC_SEC	Ordering doctor secondary specialty	Char(3)	Same coding as doctor specialty (Limited added value)	N
REV_CEN	Revenue center of ordering doctor—specific to each organization	Char(6)	May be useful for studying physician site/specialty ordering patterns but overly complicated	N
SAMP_NUM	Sample number—accession number for logging specimens	Char(6)	Numbers are not unique as they are re-used within a relatively short time period	N
TEST_NUM	Test number—instrument sequence for tracking workload within the lab	Char(6)	Used for internal management processes	N
TEST_SEQ	Test sequence – assigned by equipment in the lab that is used to run the lab test—individual transmission sequence is assigned to each test within a battery to relate results back to appropriate test	Small integer	Not useful beyond the time of original lab analysis	N
TESTTYPE	Test type—whether test is a single test or part of a battery T = single test B = battery	Char(1)		N

(Table continues)

Table 5 (continued).

Variable Name	Description	Format	Comments	Include in Lab VDW (Y/N)
	Laboratory Test Reference Sub-Table			
TESTCODE	Lab test code—relates to lab test description in reference table	Char(6)		Y
TESTDESC	LAB test description	Char(30)		Y
TESTTYPE	Test type—designates whether lab is a single test or part of a battery T = single test B = Battery	Char(1)		N
SERV_CD	Service code	Int(4)		Y
DEPTCODE	Lab department code—where test is usually performed	Char(4)	Limited usefulness	N
COL_IND	Column indicator—refers to transmission of results—only used for internal lab processing	Char(1)		N

*This file is under construction as of the time of this writing.

to grant when they understand the confidentiality protections built into the Cancer Counter.

The Cancer Counter has proven to be useful for estimating recruitment yields for new cancer research proposals, rapidly independently verifying the number of cancer cases pulled from claims and encounter data, and describing the prevalence and incidence of cancer across CRN health plans and across different demographic subgroups among their memberships.

Future CRN work will likely focus on developing additional data query tools: a Dispensing Counter to compute exposure rates to specific drug entities or classes of medications, a Cancer Prevention Counter to compute use rates of cancer prevention services (e.g., Pap smears, mammography, flexible sigmoidoscopy) in a defined population cohort, a Surgical Procedure Counter to compute use rates of surgical procedures, a Laboratory Procedure Counter, and an Imaging Procedure Counter. By building in minimum data aggregation requirements, we can provide greater access to utilization information without meaningful increases in risk to privacy of our members.

Natural Language Processing

Up to this point, we have focused on coded quantitative data. The CRN is also working on informatics tools to access and analyze free text strings in EMRs. Providers write notes in the text fields of the EMR to document clinical assessment findings, medical history, patient concerns, communications to patients and family members, care plan, and so on. In some EMRs, these text fields are exported from the online system to an end-user file that collects all the information on each case during the time interval specified by the extraction program (day, week, month, year). Traditionally, qualitative analysis software tools would be used to examine text inputs.

CRN informaticists, as part of the CRN project "Using Electronic Medical Records to Measure and Improve Adherence to Tobacco Treatment Guidelines in Primary Care" (Victor Stevens, PhD, Principal Investigator) are developing a natural language processing (NLP) tool called MediClass to determine the extent to which physicians followed the AHRQ 5A's guidelines *(28)* (i.e., assessed tobacco use, gave advice to quit, assessed readiness to change, assisted patients to quit smoking, and arranged follow-up after a quit attempt). Free-text chart notes contain information that is essential to evaluating the quality of advice and counseling activities. MediClass *(29–31)* is a generic tool that could be adapted to analysis of physician documentation of other cancer prevention services, cancer screening procedures, diagnostic work-up, treatment decision-making, palliative care choices, and end-of-life care for cancer patients. Investigators must create their library of concepts in approaching a particular analysis goal, and MediClass assists the investigators in identifying the variety of depictions of synonyms for each concept. MediClass opens up large volumes of digital text information for analysis.

Using Standardized Data for Cancer Interventions

Standardized data can serve as a foundation for cancer interventions with patients or providers. As shown in Figure 2, coded and text EMR data can serve as the foundation for interventions to improve quality of cancer care. These could be as simple as generating automated reminders to physicians when their patients need mammographies, PAP smears, or other cancer-screening procedures. These reminders could be in the form of automated e-mail messages, postcards, and oral messages during office visits. The EMR could also contain algorithms to guide physicians in selecting appropriate chemotherapy and radiation doses and cycles. Standardized files from the EMRs could be used to conduct preliminary assessments of eligibility for open clinical trials. Standardized files could serve as the foundation for patient interventions—self-management of risky behaviors (smoking, drinking, diet, inactivity), symptom self-management, self-management of medication side effects, and so on.

CRN Research Management Web Site

The CRN data standards are promoted and documented on the CRN secure Web site. The CRN Web site contains documentation (including data specifications and programs that use standardized data), the Cancer Counter, and other applications. The Web site also provides for threaded discussion groups; distribution of draft proposals, protocols, measurement instruments, research manuscripts, late-breaking news announcements, meeting agendas and minutes; and other CRN administrative information. The Web site is an essential tool to supporting the virtual research

Table 6. CRN standardized utilization data structures

Variable Name	Variable Definition	Values	Comments
	A. Utilization Data Structure*		
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
ENCTYPE	Encounter Type	Char (1)	I = Inpatient, A= Ambulatory visit, T = Telephone, E = Email, O = Other
ENCOUNTER_SUBTYPE	Encounter Subtype	Char (1)	I = Inpatient, 1 = Short Stay, 2 = Hospital Ambulatory, 3 = Hospice, 4 = Home Health, 5 = SNF, 6 = ICF, 7 = Nursing Home, 8 = Rehab, 9 = Dialysis, A = Other non-hospital
PROVIDER	Identifier unique to a provider	Character. Unique to each HMO	Physician or other provider code. Use Same coding scheme as RXMD in RX table.
ADATE	Outpatient encounter date or admit date	Numeric (4)	SAS date
DDATE†	Discharge date	Numeric (4)	SAS date; missing for outpatient visit
DISCHARGE_DISPOSITION†	Discharge status	Char (1)	A = Discharged alive, E = Expired, U = Unknown, Blank for outpatient visit
DEPARTMENT	Department Code (specialty providing service)	Char	Optional. Outpatient only. See codes below
ADMITING_SOURCE†	Admitting Source	Char	Optional. CLIN = Outpatient Clinic, ER = Emergency Room, HOSP = Transfer from Another Hospital, SNF = Skilled Nursing Facility, ICF = Intermediate Care Facility, HH = Home Health, HOSPICE = Hospice, RES = Residential Facility, OTH = Other, UNK = Unknown
FACILITY_CODE	Facility code that identifies hospital or clinic	Char (12)	Optional.
DISCHARGE_STATUS†	Hospital Discharge Status	Char (2)	Optional. Coding according to http://tinyurl.com/2bnud or http://www.hce.org/Medicare/Word_Documents/Patient_Discharge_Documents/Pt_Discharge_Status_Codes.doc
DRG†	Diagnostic Related Group	Char (3)	Optional. Data quality issues. Blank for outpatient visit
	B. Diagnosis Data Sub-structure		
A record is a diagnosis code unique to an index variable (MRN, ENCTYPE, PROVIDER, ADATE) combination.			
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
ENCTYPE	Encounter Type	Char (1)	I = Inpatient, A= Ambulatory visit, O = Other
PROVIDER	Identifier unique to a provider	Char (6)	Physician or other provider code
ADATE	Outpatient encounter date or admit date	Numeric (4)	SAS date
DX‡	ICD-9-CM diagnosis codes	Char (6)	Note the decimal point. Clean up site introduced suffixes; xxx.xx, Vxx.xx, Exxx.x
PDX	Principal DX flag	Char (1)	P = principal, S = secondary, X = not available
	C. Procedure Data Sub-structure		
MRN	Identifier unique to an individual	Character. Unique to each HMO	Used to link across files
ENCTYPE	Encounter Type	Char (1)	I = Inpatient, A= Ambulatory visit, O = Other
PROVIDER	Identifier unique to a provider	Char (6)	Physician or other provider code
ADATE	Outpatient encounter date or admit date	Numeric (4)	SAS date
PX§	Procedure code	Char (6)	xx.xx = ICD-9-CM, xxxxx = CPT-4 or any other code.
CODETYPE	Code type flag	Char (1)	I = ICD-9-CM, C = CPT- 4, H = HCPCS, L = local homegrown, O = Other

(Table continues)

Table 6 (continued).

Variable Name	Variable Definition	Values	Comments
	D. Provider Specialty Data Sub-structure		
PROVIDER	Identifier unique to an individual	Character. Unique to each HMO	
SPECIALTY	Specialty code	Char (3)	Using specialty coding system shown below:
Specialty Code	Specialty Description	Specialty Code	Specialty Description
ACUP	Acupuncture	IM	Internal Medicine
ALGY	Allergy	IMUN	Immunology
AMBU	Ambulance Services	IND	Industrial Medicine
ANES	Anesthesiology	INF	Infectious Disease
AUD	Audiology	IR	Injection Room
CARD	Cardiology	LAB	Laboratory
CASR	Cast Room	MH	Mental Health
CHEM	Chemical and Alcohol Dependency	NATU	Naturopathy
CHIR	Chiropractic	NEPH	Nephrology
CMHL	Community Health	NEUR	Neurology
CRIT	Critical Care Medicine	NEWB	Newborn
CRMG	Care Management	NRSG	Neurosurgery
DENT	Dental	NUCL	Nuclear Medicine
DERM	Dermatology	NUT	Nutrition
EDUC	Education	OBGN	Obstetrics/Gynecology
ENDO	Endocrinology	OCTH	Occupational Therapy
ENT	Otolaryngology	ONC	Oncology
ER	Emergency Room	OPTH	Ophthalmology
FP	Family Practice	OPTO	Optometry
GEN	Genetics	ORTH	Orthopedics
GER	Gerontology/Geriatrics	OST	Osteopathy
GI	Gastro-Intestinal Medicine	PATH	Pathology
HAP	Health Appraisals	PEDS	Pediatrics
HEP	Hepatology	PERI	Perinatology
HOSP	Hospital Care	PHYS	Physiatry
POD	Podiatry	RN	Registered Nurse
PSRG	Plastic Surgery	SPOR	Sports Medicine
PT	Physical Therapy	SPTH	Speech Therapy
PULM	Pulmonary Medicine	SURG	General Surgery
RAD	Radiology	URG	Urgent Care
REHB	Rehabilitation	URO	Urology
RESP	Respiratory Therapy	OTH	Other
RHEU	Rheumatology	UNK	Unknown

*The utilization data structure consists of four tables: 1) UTILIZATION: characterizes the outpatient visit or hospital say; 2) DIAG: the diagnosis codes associated with the Utilization record; 3) PROC: the procedure codes associated with the Utilization record; and 4) SPEC: lookup table for provider specialty. The linking variables are MRN, ENCTYPE, PROVIDER, and ADATE. A unique combination of these variables is an individual outpatient visit or an inpatient stay. A single visit or a single hospital stay will have a single record in the UTILIZATION file. Each DX code at a visit or a stay will have a separate record in the DIAGNOSIS data table. Each procedure code at a visit or a stay will have a separate record in the procedure data table. The PROVIDER variable is most useful for an outpatient visit. For an inpatient stay the major goal is that PROVIDER be consistent within an HMO. If possible, use the admitting physician. An inpatient stay has a single PROVIDER, even if multiple providers performed procedures.

†Inpatient Stays only.

‡Exclude rule outs if possible. Include denied claims if you consider the utilization to be valid.

§Exclude rule outs if possible. Include denied claims if you consider the utilization to be valid.

organization, along with e-mail, conference calls, and semi-annual meetings.

Results

Current Status

Content areas for which data dictionaries have been developed include cancer registry, demographic attributes, enrollment periods, pharmacy, utilization (including diagnoses and procedures), and laboratory procedures. Programming and testing of these data at this writing has varied across sites but is largely complete. Three sites do not maintain or have access to cancer registries, but the other sites have standardized tumor registry data for dates beginning anywhere from 1960 to 1995. Ten sites have standardized the enrollment data available, nine have standardized pharmacy data, and seven have standardized utilization data (see Table 7).

To date, standardized files have been used extensively to determine the feasibility of conducting studies, as well as to generate enrollment and utilization data for a study of racial disparities in cancer survival. Data requests are generally accompanied by a SAS program (written by a programmer or SDM) that uses the standardized data elements to create the requested output. With the data already in specified formats, requests typically require minimal effort on the part of other site programmers.

Several SDMs have begun to write macros to use in conjunction with standardized data files. These can be used across sites for common requests, such as identifying health plan members who are continuously enrolled for a given period of time,

Table 7. CRN standardized data availability by year*

Type of Data	Health Plans										
	Fallon/MPCI	GHC	HFHS/HAP	HPRF	HPHC	KPCO	KPGA	KPH	KPNC	KPNW	KPSC
Enrollment Tracking	1987	1988	1980	1990	1969	Aug 1993	1995	1958	1980	1982	1988
Ambulatory Visits	1987	1992	1988	1990	1969	1994	1995	1989	1995	1987 (KARE) 1998 (EPIC)	1992
Hospital Use	1987	1979	1989	1990	1990	1994	1995	1987	1979	1965	1990
Pharmacy	1987	1977	1992	1990	1988	1992	1995	1987	1993	1986	1992
Imaging	1996	1986	1988 partial	1990 text	1969 text	1992	1995	1988	1992	1988	—
Laboratory	1990	—	1995	1994	1969	1992	1995	1988	1994	1993	1990
Home Health	—	—	1995	1990	1990	1994	1995	—	1995	1987	—
Hospice			1995						1995	1987	1994
Nursing Home			1995						1995	1987	
Tumor Registry	1973	1974	1972	1974	1982	1987	1995	1973	1973	1960	1988
Claims	1987	1979	1991	1990	1990	1994	1995	1995	1991	1987	1991
Vital Statistics	2002	1972	2005	1981	?	2000	?	?	1970	1990	1975
Durable Med. Equip.			1995						1996	1980	

*— = information not available in a single comprehensive data file. "Text" = free text data with search capability. "KARE" = electronic appointments and registration system. "EPIC" = electronic medical record system. Fallon/MPCI = Fallon Health Systems/Meyers Primary Care Institute. GHC = Group Health Cooperative. HFHS/HAP = Henry Ford Health System/Health Alliance Plan. HPRF = HealthPartners Research Foundation. HPHC = Harvard Pilgrim Health Care. KPCO = Kaiser Permanente Colorado. KPGA = Kaiser Permanente Georgia. KPH = Kaiser Permanente Hawaii. KPNC = Kaiser Permanente Northern California. KPNW = Kaiser Permanente Northwest. KPSC = Kaiser Permanente Southern California.

identifying patients who were prescribed a certain drug, or computing the Charlson-Deyo comorbidity index.

Research Project Access to Standardized Data

If an investigator has an interest in developing a research proposal using CRN data, he or she should begin with informal discussions with one of the members of the CRN Steering Committee or the NCI Project Officer for CRN. The CRN New Proposals Committee must review all research proposal concepts. If a CRN internal collaborator can be identified, then the steps preparatory to research can begin. The instructions and forms for this process are downloadable from the CRN secure Web site (thus the need for an internal sponsor/collaborator). The Web site contains much useful information for planning a research project, as well as creating tables to include in a proposal.

Single-site proposals do not need to go through the CRN; we are focusing strictly on research proposals that require data from two or more participating health plans. As the proposal is completed, the chief research officer or other appropriate executive-level official at each participating health plan must approve it.

In the human subjects section of a proposal, the provisions for protecting privacy, confidentiality, and data security must be well defined. All CRN research projects must be in the public domain; no proprietary confidential research can be conducted with CRN resources.

Potential users of the standardized CRN data should proceed by seeking permission to view the CRN secure Web site. New projects should select collaborating sites and then file the requisite HIPAA forms for each site. This step allows data extraction and analysis for proposal writing, before IRB approval is obtained for the funded project. SAS programs are distributed to each site to extract proposal data. The extracted data may be in the form of tables of anonymous data or limited data sets. These are sent to the central site for merging and inclusion in the proposal. Once the proposal is funded, application is made to the relevant IRBs and HIPAA privacy and data security officers to approve the proposed data extraction, transfer, storage, and analysis procedures. For secondary data, SAS data extract files are again sent out to the local sites to run against the standardized local files, and limited data sets are transferred via secure encrypted Web transfer to the coordinating center site to create analysis files.

The CRN Publications Committee reviews all manuscripts before submission to ensure that they accurately state CRN's role; the committee also may comment on the scientific substance of the manuscript.

Cancer Biomedical Informatics Grid (caBIG)

The CRN was launched in 1998, before the NIH Roadmap initiative and the caBIG component of the Roadmap (http://caBIG.nci.nih.gov/). CaBIG is an informatics infrastructure that will connect teams of cancer and biomedical researchers to enable them to better develop and share tools and data in an open environment with common standards. CaBIG will create a voluntary virtual network (i.e., "grid") that links individuals and institutions both nationally and internationally, effectively forming a World Wide Web of cancer research. The caBIG pilot program was launched in 2003 with more than 50 NCI-designated cancer centers working in partnership with NCI to develop the vision, approach, and structure of caBIG.

To date, CRN and caBIG have evolved independently, but in parallel fashion. CRN is committed to participating in caBIG and working toward unification of CRN information resources with the caBIG. Of the many significant challenges to be worked out in this process, two important ones include developing a bridge between licensed proprietary operating systems and database and data analysis software on one hand and open source operating systems and data software on the other, and converting the variety of data standards that currently govern health plan data to national data standards (e.g., HL7, SNOMED CT, UMLS, LOINC), particularly standardized medical vocabularies and common data elements.

Conclusions

The CRN, a collaborative virtual research organization, performs the following key functions to facilitate multisite

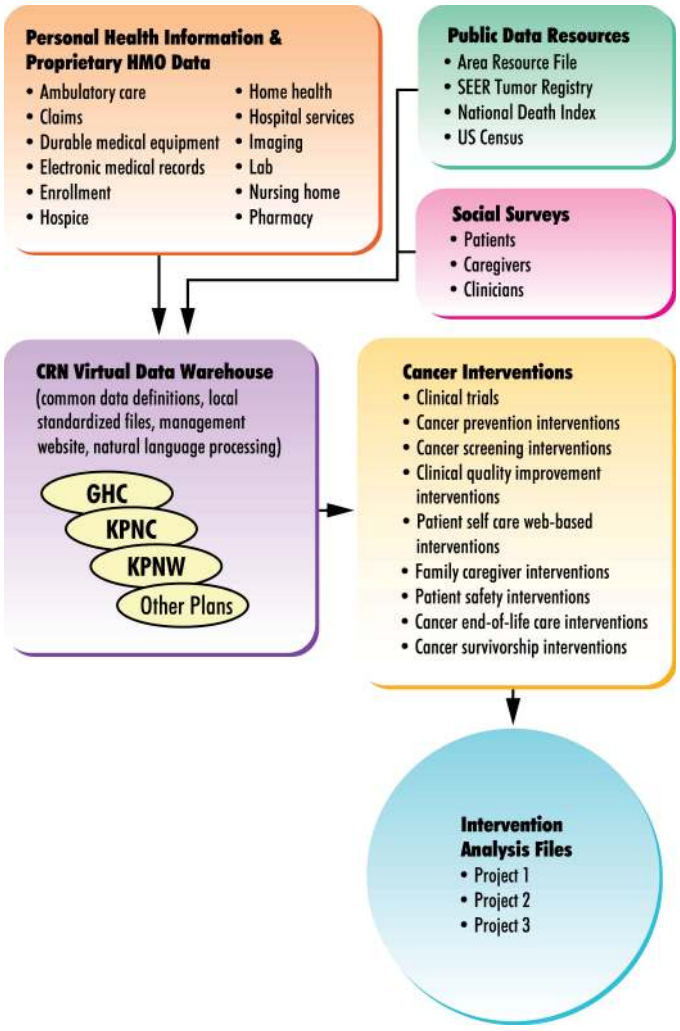

Fig. 2. Using the CRN standardized files for cancer interventions.

collaborative research: a governance structure and operating policies that provide a first point of contact for interested researchers and referral to potential collaborators (both sites and investigators); a comprehensive user-friendly view of health plan data resources and the CRN's standardized data files through its Web site; data query tools that provide frequencies and cross-tabulations on incidence of tumors for each health plan and for any combination of health plans; encrypted transfer of research data through its secure Web site; support for proposal-writing groups, paper-writing groups, project teams, and individual researchers and research staff by providing an information clearinghouse, threaded discussion groups, contact information for all CRN staff members, and document version control; and an archive for project documentation, CRN policies and procedures, presentations, survey research instruments, medical record abstraction forms and coding instructions, and other items. These functions are necessary support for the national clinical trials and coordinate studies network envisioned in the NIH Roadmap. The CRN sites are participating in a contract from the National Heart Lung and Blood Institute to develop a standardized clinical trials infrastructure in our health plans and research centers.

Clinical and business data captured in the course of health care encounters is a highly valuable asset for clinicians, health plans, payers, policy makers, and researchers. Researchers who are employees of integrated delivery systems have access to databases of personal health information on their health plan's members and possess high levels of local knowledge of health plan data and delivery systems. HIPAA regulations define health plans as covered entities. The CRN, by virtue of the composition of the memberships of the participating health plans, can provide a large database of cancer patients of all ages. Comparison between SEER-Medicare cases and CRN Medicare cases enables research on the effects of different health care arrangements—integrated delivery systems versus independent fee-for-service indemnity practice—on patterns of cancer diagnosis, treatment, and palliation. Thus, CRN represents a data resource with high relevance to cancer research and policy.

Even with the HIPAA, IRB, and proprietary business constraints, the CRN standardized data are a quasi-public resource. Their use is strongly regulated to protect the privacy rights of the individuals whose data are included in these systems and to protect the commercial interests of the participating health plans and medical groups. CRN stands ready to collaborate with researchers from outside institutions in developing and conducting innovative public domain research. Moreover, the CRN is mapping out a multifaceted strategy to connect with caBIG and work toward convergence of CRN data standards with the caBIG common data standards.

The CRN provides rapid, systematic, efficient access to key research leaders, investigators, and health plan managers who may be interested in and affected by a potential research project. Moreover, researchers viewing the standardized file specifications may find potential research topics from the intersection of variable sets across the standardized data files. The majority of CRN health plans have EMRs, which provide rich and accessible data sets.

Perhaps the most important CRN function is maintaining access to the research laboratories by working with health plan executives and managers to mediate conflicts between the priorities in clinic operations and those in research projects. Future developments could include standardized files on clinicians and facilities, as well as expanding the variables included on the utilization files to include chemotherapy, hormone therapy, and radiotherapy for cancer patients, as well as home health, hospice, and end-of-life care.

References

(1) Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Mariotto A, Feuer EJ, Edwards BK (eds.). SEER Cancer Statistics Review, 1975–2001. Bethesda (MD): National Cancer Institute. Available at: http://seer.cancer.gov/csr/1975_2001/index.html.

(2) Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 2002;40:3–18.

(3) Vogt TM, Lafata JE, Tolsma D, Greene SM. The role of research in integrated health care systems: the HMO Research Network. Am J Manag Care 2004;10:643–8.

(4) Wagner EH, Greene SM, Hart G, Field TS, Fletcher S, Geiger AM, Herrinton LJ, Hornbrook MC, Johnson CC, Mouchawar J, Rolnick SJ, Stevens VJ, Taplin SH, Tolsma D, Vogt TM. Building a research consortium of large health systems: the Cancer Research Network. J Natl Cancer Inst Monogr 2005;35:3–11.

(5) Chan KA, Davis RL, Gunter MJ, Gurwitz JH, Herrinton LJ, Nelson WW, Raebel MA, Roblin DW, Smith DH, Platt R. The HMO Research Network. Chapter 18. Strom B. Pharmacoepidemiology (4th ed.).

(6) Rigotti NA, Quinn VP, Stevens VJ, Solberg LI, Hollis JF, Rosenthal AC, Zapka JG, France E, Gordon N, Smith S, Monroe M. Tobacco-control

policies in 11 leading managed care organizations: progress and challenges. Eff Clin Pract 2002;5:130–6.
(7) Quinn VP, Stevens VJ, Hollis JF, Rigotti NA, Solberg LI, Aickin M, et al. Tabacco-cessation services and patient satisfaction in nine nonprofit HMOs. Amer J Preven Med 2005;29:77–84.
(8) Solberg LI, Hollis JA, Stevens VJ, Rigotti NA, Quinn VP, Aickin M. Does methodology affect the ability to monitor tobacco control activities? Implications for HEDIS and other performance measures. Prev Med 2003;37:33–40.
(9) Solberg LI, Quinn VP, Stevens VJ, Vogt TM, Rigotti NA, Zapka JG, Ritzwoller DP, Smith KS. Tobacco control efforts in managed care: what do the doctors think? Am J Manag Care 2004;10:193–8.
(10) Taplin SH, Ichikawa L, Yood MU, Manos MM, Geiger AM, Weinmann S, Gilbert J, Mouchawar J, Leyden WA, Altaras R, Beverly RK, Casso D, Westbrook EO, Bischoff K, Zapka JG, Barlow WE. Reasons for late-stage breast cancer among women with access to care: Absence of screening, absence of detection, or potential breakdown in followup? J Nat Cancer Inst 2004;96:1518–27.
(11) Mouchawar J, Valentine Goins K, Somkin C, Puleo E, Hensley Alford S, Geiger AM, Taplin S, Gilbert J, Weinmann S, Zapka J. Guidelines for breast and ovarian cancer genetic counseling referral: adoption and implementation in HMOs. Genet Med 2003;5:444–50.
(12) Goins KV, Zapka JG, Geiger AM, Solberg LI, Taplin S, Yood MU, Gilbert J, Mouchawar J, Somkin CP, Weinmann S. Implementation of systems strategies for breast and cervical cancer screening services in health maintenance organizations. Am J Manag Care 2003;9:745–55.
(13) Zapka JG, Taplin SH, Solberg LI, Manos MM. A framework for improving the quality of cancer care: the case of breast and cervical cancer screening. Cancer Epidemiol Biomarkers Prevention 2003;12:4–13.
(14) Zapka JG, Puleo E, Taplin SH, Goins KV, Ulcickas Yood M, Mouchawar J, Somkin C, Manos MM. Processes of care in cervical and breast cancer screening and follow-up—the importance of communication. Prev Med 2004;39:81–90.
(15) Buist D. Hormone therapy prescribing patterns in the United States. Obstet Gynecol 2004;104:1042–50.
(16) Field TS, Cernieux J, Buist D, Geiger A, Lamerato L, Hart G, Bachman D, Krajenta R, Greene S, Hornbrook MC, Ansell G, Herrinton L, Reed G. Retention of enrollees following a diagnosis of cancer within health maintenance organizations in the Cancer Research Network. J Natl Cancer Inst 2004;96:148–52.
(17) Geiger AM, Yu O, Herrinton LJ, Barlow WE, Harris EL, Rolnick S, Barton MB, Elmore JG, Fletcher SW. A population-based study of bilateral prophylactic mastectomy efficacy in women at elevated risk for breast cancer in community practices. Arch Intern Med 2005;165:516–20.
(18) Puleo E, Zapka J, White MJ, Mouchawar J, Somkin C, Taplin S. Caffeine, cajoling and other strategies to maximize clinician survey response rates. Eval Health Prof 2002;25:169–84.
(19) Rolnick SJ, Hart G, Barton MB, Herrinton L, Flores SK, Paulsen KJ, Husson G, Harris EL, Geiger AM, Elmore JG, Fletcher SW. Comparing breast cancer case identification using HMO computerized diagnostic data and SEER data. Am J Manag Care 2004;10:257–62.
(20) Reisch LM, Fosse JS, Beverly K, Yu O, Barlow WE, Harris EL, Rolnick S, Barton MB, Geiger AM, Herrinton LJ, Greene SM, Fletcher SW, Elmore JG. Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol 2003;157:546–51.
(21) Geiger AM, Greene SM, Pardee RE 3rd, Hart G, Herrinton LJ, Macedo AM, Rolnick S, Harris EL, Barton MB, Elmore JG, Fletcher SW. A computerized system to facilitate medical record abstraction in cancer research. Cancer Causes Control 2003;14:469–76.
(22) Field TS, Cadoret CA, Brown ML, Ford M, Greene SM, Hill D, Hornbrook MC, Meenan RT, White MJ, Zapka JM. Surveying physicians: do components of the "Total Design Approach" to optimizing survey response rates apply to physicians? Med Care 2002;40:596–606.
(23) Ford ME, Hill DD, Nerenz D, Hornbrook M, Zapka J, Meenan R, Greene S, Johnson CC. Categorizing race and ethnicity in the HMO cancer research network. Ethn Dis 2002;12:135–40.
(24) Foster I, Kesselman C, Tuecke S. The anatomy of the grid: enabling scalable virtual organization. Int J High Performance Comput Appl 2001; 15:200–22.
(25) Hornbrook MC, Goodman MJ, Fishman PA, and Meenan RT. Health-based payment and computerized patent record systems. Eff Clinical Pract 1998;1:66–72.
(26) Hornbrook MC, Goodman MJ, Fishman PA, Meenan RT, O'Keeffe-Rosetti M, and Bachman DJ. Building health plan databases to risk-adjust outcomes and payments. Int J Qual Health Care 1998;10:531–8.
(27) Fishman PA, Hornbrook MC, Meenan RT, Goodman MJ. Opportunities and challenges for measuring cost, quality, and clinical effectiveness in health care. Med Care Res Rev 2004;61:124S–43S.
(28) Fiore MC, Bailey WC, Cohen SJ, et al. Treating Tobacco Use and Dependence: A Clinical Practice Guideline. Rockville, MD: US Department of Health and Human Services; 2000. Available at: http://www.surgeongeneral.gov/tobacco/treating_tobacco_use.pdf.
(29) Hazlehurst B. MediClass: A general purpose system for clinical events detection in the EMR. HMO Research Network Annual Conference, Dearborn MI; May 3–5, 2004.
(30) Hazlehurst B. The Mediclass system for identifying vaccine reactions in the EMR. Annual Meeting of the Vaccine Safety Datalink (VSD), Madison, WI; June 2004.
(31) Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: a system for detecting and classifying encounter-based clinical events in any EMR. J Am Med Inform Assoc 2005;12:517–29.

Notes

M. C. Hornbrook holds stock in Pfizer, Inc., and is currently conducting research sponsored by that company on the economic burden of generalized anxiety disorder. The pharmaceutical company–sponsored work has no relationship to this article.

The Cancer Research Network, Grant Number U19 CA 79689 from the National Cancer Institute, supported this research. The CRN consists of the research programs, enrollee populations and databases of 11 integrated healthcare organizations that are members of the HMO Research Network. The health care delivery systems participating in the CRN are: Group Health Cooperative, Harvard Pilgrim Health Care, Henry Ford Health System/Health Alliance Plan, HealthPartners Research Foundation, the Meyers Primary Care Institute of the Fallon Healthcare System/University of Massachusetts, and Kaiser Permanente in six regions: Colorado, Georgia, Hawaii, Northwest (Oregon and Washington), Northern California and Southern California. The 11 health plans, with nearly ten million enrollees, are distinguished by their long-standing commitment to prevention and research, and collaboration among themselves and with affiliated academic institutions.

The CRN site principal investigators are Thomas M. Vogt, MD, MPH (Kaiser Permanente Hawaii Region, Lisa Herrinton, PhD (Kaiser Permanente Northern California Region), Ann Geiger, PhD (Kaiser Permanente Southern California Region), Mark C. Hornbrook, PhD (Kaiser Permanente Northwest Region), Edward H. Wagner, MD, MPH (overall Principal Investigator for the CRN) (Group Health Cooperative), Judy Mouchawar, MD, MSPH (Kaiser Permanente Colorado Region), Cheri Rolnick, PhD (HealthPartners Research Foundation), Christine Cole Johnson, PhD (Henry Ford Health System), Suzanne Fletcher, MD (Harvard Pilgrim Health Care), Dennis Tolsma, MPH (Kaiser Permanente Georgia Region), Terry Field, ScD (Fallon Health System/Meyer Primary Care Institute), and Margaret Gunter, PhD (Lovelace Clinic Foundation). The CRN Site Data Managers are Jennifer Ellis (KPCO), Gene Hart (GHC), Mark Schmidt (KPH), Irina Miroshnik (HPHC), Don Bachman (KPNW), Liyan Liu (KPNC), Ann Geiger (KPSC), Lois Lamerato (HFHS), Karen Wells (HFHS), Rob Diseker (KPGA), Amy Butani (HPRF), and Hassan Fouayzi (Meyers/Fallon).

Martha Swain and Margaret Sucec provided technical editorial assistance.