Feasibility of Linking Population-Based Cancer Registries and Cancer Center Biorepositories

Margaret E. McCusker,¹ Rosemary D. Cress,^{2,3} Mark Allen,²
Allyn Fernandez-Ami,² and Regina Gandour-Edwards⁴

Purpose: Biospecimen-based research offers tremendous promise as a way to increase understanding of the molecular epidemiology of cancers. Population-based cancer registries can augment this research by providing more clinical detail and long-term follow-up information than is typically available from biospecimen annotations. In order to demonstrate the feasibility of this concept, we performed a pilot linkage between the California Cancer Registry (CCR) and the University of California, Davis Cancer Center Biorepository (UCD CCB) databases to determine if we could identify patients with records in both databases. **Methods:** We performed a probabilistic data linkage between 2180 UCD CCB biospecimen records collected during the years 2005–2009 and all CCR records for cancers diagnosed from 1988–2009 based on standard data linkage procedures. **Results:** The 1040 UCD records with a unique medical record number, tissue site, and pathology date were linked to 3.3 million CCR records. Of these, 844 (81.2%) were identified in both databases. Overall, record matches were highest (100%) for cancers of the cervix and testis/other male genital system organs. For the most common cancers, matches were highest for cancers of the lung and respiratory system (93%), breast (91.7%), and colon and rectum (89.5%), and lower for prostate (72.9%). **Conclusions:** This pilot linkage demonstrated that information on existing biospecimens from a cancer center biorepository can be linked successfully to cancer registry data. Linkages between existing biorepositories and cancer registries can foster productive collaborations and provide a foundation for virtual biorepository networks to support population-based biospecimen research.

Introduction

BIOspecimens ARE VITAL CANCER research tools that can be used to evaluate tumor molecular profiles, describe the molecular epidemiology of newly identified oncogenes and their impact on recurrence and survival, and identify biomarkers that predict which patients are most likely to benefit from specific therapies.^{1,2} This research can be enhanced by linking information on existing biospecimens to population-based cancer registry data to create population-based biorepositories. Population-based cancer registries collect information on all cancer cases diagnosed within their catchment area from hospitals, pathology laboratories, physician offices, and radiation therapy facilities. The information includes patient demographics, primary tumor characteristics, first course of cancer treatment, and long-

term follow-up. Cancer registries monitor survival through linkages to data sources such as motor vehicle and voter registration files, and to death certificates.

Population-based biorepositories can provide data and biospecimens that are representative of their underlying population to minimize selection bias and supply information about patients for whom biospecimens are not available, which can be used to validate research findings. This combination of detailed, representative data and biospecimens for a specific population can be used to assess the generalizability of findings from smaller research studies and to describe how biomarkers are expressed at a population level. Biospecimens from population-based biorepositories can also be used to study the molecular epidemiology of rare tumors and tumors among specific population subgroups for which sufficient information may be difficult to assemble.^{3–5}

¹Cancer Surveillance and Research Branch, California Department of Public Health, Sacramento, California, at the time of writing this article.

²Public Health Institute, California Cancer Registry, Sacramento, California.

³Department of Public Health Sciences, University of California, Davis, California.

⁴Cancer Center Biorepository, Department of Pathology and Laboratory Medicine, University of California Davis Medical Center, Sacramento, California.

Because biorepositories are expensive to establish and maintain, we wanted to test the feasibility of creating a “virtual” cancer registry-linked biorepository, with information on previously collected biospecimens from existing biorepositories linked to population-based cancer registry data. We conducted a test linkage between the California Cancer Registry (CCR) and the University of California, Davis Cancer Center Biorepository (UCD CCB) databases to determine if the data could be matched to identify patients with records in both databases. Since the UCD CCB biorepository data are likely to be similar to data in other cancer center-based biorepositories, we hypothesized that this test linkage would provide critical information on the feasibility of creating a virtual biorepository. These linkages could allow researchers to use the detailed demographic, clinical and follow-up information available from the cancer registry in conjunction with the biospecimens available from the biorepository to conduct population-based biospecimen research.

Materials and Methods

The University of California, Davis (UCD) Institutional Review Board (IRB) approved the study protocol. In order to protect patient confidentiality, the linkage was conducted behind CCR firewalls, and results are reported in aggregate.

The CCR is California’s statewide population-based cancer surveillance system, administered by the California Department of Public Health, which has collected data on incident cancers diagnosed among California residents since 1988. Data are collected through a network of regional registries, which also participate in the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI). For each cancer patient, CCR collects demographic, tumor, treatment and survival information.

The UCD CCB, a UCD Comprehensive Cancer Center Shared Resource funded by NCI and the UCD Department of Pathology and Laboratory Medicine, has collected tissue, blood, and other biological material from patients who receive care at UCD Medical Center under strict IRB-approved protocols since 2004. Tissue specimens that are considered remainder or “leftover” after pathologic evaluation are obtained from routine diagnostic or therapeutic surgical procedures. Blood specimens are procured as an “extra” specimen during scheduled phlebotomy services. Specimens are obtained with full patient consent and distributed to researchers with non-identifiable clinical information in a modified “honest broker” model.⁶ Patients choose whether their stored biospecimens may be used for specific cancer research projects or for research about other health problems.

A list of UCD CCB biospecimen records was extracted from the CCB database (caTissue), converted to an Excel file, and sent to CCR. The list included the following data elements for each biospecimen: first name, middle initial, last name, gender, date of birth, race/ethnicity, medical record number, tissue site, pathological status (e.g., benign, malignant), pathology specimen date, and surgical pathology number. Most patients who donated biospecimens to the UCD CCB had more than one specimen in the biorepository, and some had more than one specimen stored on the same date. To streamline this linkage and ensure that only one record per patient for each date was used, only records with unique values for medical record number, tissue site, and pathology specimen date were used. New elements were

created for the UCD data that recoded race/ethnicity, tumor site, and pathological status to match the standard cancer registry codes used by CCR. Information on pathological status in the UCD CCB clinical annotation was recoded as follows to align with the tumor behavior codes defined in the International Classification of Diseases for Oncology, Third Edition: benign, in situ, malignant, and unknown.⁷

The following CCR data elements were used for the linkage: first name, last name, maiden name, gender, date of birth, race/ethnicity, medical record number, tissue site, tumor behavior, date of diagnosis (proxy for pathology report date), pathology report number, a flag to indicate that the patient had received care at UCD, and a created flag to indicate whether the CCR address at diagnosis was located within the UCD medical center catchment area. For the CCR data, all cancers diagnosed for all individuals were used in the linkage. Thus, if a person had more than one cancer diagnosed, all of the tumors were included for potential linkage. At the time of the linkage, the CCR data were considered complete (at least 95% of expected cases reported to CCR) for cases diagnosed through 2008, and were almost 90 percent complete for cases diagnosed during 2009.

The linkage between the databases was based on the standard CCR probabilistic data linkage procedures (http://www.ccr.ca.gov/Data_and_Statistics/Cancer_Data_for_Research.shtml). The process comprised six sequential comparisons of the two data sets, which accounted for possible differences in how data elements were recorded, such as typographical errors or variations in coding from the medical record that were not true differences. During the linkage, certain data elements were required to be exact matches and the remaining data elements were used to create an agreement weight for a pair of records. Data elements with the same value in the UCD CCB and CCR records received a positive agreement weight. If the values were different, the data element received a negative weight. The weights for all of the data elements were added, and a histogram that depicted the frequency distribution of the weights was manually reviewed to determine high and low cut-off values for matches. Records with high total weights were considered matches. Records with medium total weights were individually reviewed and a determination based on examination of all data element values was made to classify the pair as a match or a nonmatch. Records with low total weights were classified as nonmatches. For each comparison, all CCR records were used so that patients with more than one distinct tumor could be matched for each tumor. Thus, if a patient had two specimens from two separate occasions in the UCD CCB database, both specimens would be counted as matches. We reviewed the results of the linkage and determined the proportions and means where appropriate.

Results

The UCD CCB databases provided 2180 records for patients diagnosed with cancer during 2005–2009 to CCR for the linkage. Of these, 1040 records had a unique value for medical record number, tissue site, and pathology specimen date. Individuals had a median of two biospecimen records (range 1–36 records). The 1040 UCD CCB records were compared with 3.3 million CCR records. A total of 844 (81.2%) of the 1040 records were matched between the UCD CCB and CCR databases. Agreement was highest for 2008,

TABLE 1. NUMBER AND PROPORTION OF CANCER RECORDS MATCHED IN THE CALIFORNIA CANCER REGISTRY (CCR) AND UC DAVIS CANCER CENTER BIOREPOSITORY (UCD CCB) DATABASES BY YEAR, 2005–2009

Year	Number of Records Used	Number of Matches	% Matches
2005	73	45	61.6%
2006	422	330	78.2%
2007	173	158	91.3%
2008	93	87	93.5%
2009	279	224	80.3%
Total	1040	844	81.2%

with 87 of 93 records (93.5%) matched, and lowest for 2005, with 45 of 73 (61.6%) records matched (Table 1). Matches for 2009 (80.3%) were slightly lower than for 2008.

Table 2 shows the proportion of records in the linkage that had identical information in both the UCD CCB and the CCR databases by data element. Agreement was greatest for gender, for which 839 (99.4%) records had the same value. Agreement was lowest for tumor behavior, for which 361 (42.8%) records had the same value.

Table 3 compares the race/ethnicity classification recorded in the two databases for all matched records. Overall, 545 of 844 records (64.6%) had the same race classification in both databases. Of 299 records with different race classifications, 254 had an unknown race in the UCD CCB database; of these, 251 (98.8%) had a race coded in the CCR database. Most of the records in the UCD CCB database without a coded race were classified as white in the CCR database (226 of 254, or 88.9%). There were 45 records where a known race classification differed between the databases (i.e., classified as white in one database and black in the other).

For Hispanic ethnicity, 689 of 844 records (81.6%) had the same ethnicity classification recorded in both databases. Of 155 records with different classifications, there were 116 with an unknown Hispanic ethnicity classification in the UCD CCB database, and four in the CCR database. There were 39 records for which a known Hispanic ethnicity classification differed between the two databases.

Table 4 shows the number of records in the linkage that were identified in both databases by cancer site. For the most common cancers, matches were highest for cancers of the lung and respiratory system (93%), breast (91.7%), and colon and rectum (89.5%), and lowest for cancers of the prostate (72.9%).

TABLE 2. AGREEMENT BETWEEN DATA ELEMENTS IN THE CALIFORNIA CANCER REGISTRY (CCR) AND UC DAVIS CANCER CENTER BIOREPOSITORY (UCD CCB) DATABASES, 2005–2009 ($n = 844$)

Variable	Number in Agreement	% Agreement
Cancer site code	724	85.8%
Tumor Behavior	361	42.8%
First name	812	96.2%
Last name	927	98.0%
Date of birth	546	64.7%
Gender	839	99.4%
Medical record number	707	83.8%
Pathology report number	478	56.6%

TABLE 3. COMPARISON OF RACE AND ETHNICITY CLASSIFICATION BETWEEN RECORDS MATCHED BETWEEN THE CALIFORNIA CANCER REGISTRY (CCR) AND UC DAVIS CANCER CENTER BIOREPOSITORY (UCD CCB) DATABASES, 2005–2009 ($n = 844$)

Race/Ethnicity	UCD CCB Classification (%)	CCR Classification (%)
White	534 (63.3)	749 (88.7)
Black	26 (3.1)	51 (6.0)
American Indian or Alaskan Native	6 (0.7)	3 (0.4)
Asian or Pacific Islander	24 (2.8)	28 (3.3)
Other or unknown	254 (30.1)	13 (1.5)
Hispanic ethnicity	55 (6.5)	57 (6.8)

Discussion

This pilot linkage demonstrated that existing cancer center biorepository data can be linked successfully to cancer registry data. As a result of this linkage, over 800 biospecimens were matched to their corresponding cancer registry data. Because the specimens were linked to previously-collected, high-quality cancer registry data from a population-based, state-wide cancer registry, information on patient demographics, tumor characteristics, and first course of treatment is readily available to supplement biospecimen data. The

TABLE 4. RECORD MATCHES BY CANCER SITE IN THE CCR AND UCD CCB DATABASES, 2005–2009 ($n = 1040$)

Description	Number of Records Used	Number of Matches	% Matches
Cervix	5	5	100.0%
Testis/other male genital system	7	7	100.0%
Corpus and uterus, NOS	37	36	97.3%
Respiratory system	114	106	93.0%
Breast	48	44	91.7%
Kidney	108	97	89.8%
Bladder	49	44	89.8%
Colorectal	38	34	89.5%
Other urinary system	9	8	88.9%
Lymphoma	18	16	88.9%
Stomach	8	7	87.5%
Pancreas	40	34	85.0%
Endocrine system	12	10	83.3%
Brain and other nervous system	23	19	82.6%
Other digestive system	10	8	80.0%
Liver	10	8	80.0%
Ovary	33	25	75.8%
Prostate	376	274	72.9%
Oral cavity and pharynx	18	13	72.2%
Soft tissue	48	34	70.8%
including heart			
Miscellaneous	18	12	66.7%
Bones and joints	6	2	33.3%
Skin	5	1	20.0%
Total	1040	844	81.2%

NOS, not otherwise specified.

number of specimens for specific sites was small, but could be increased substantially by expanding this linkage to other cancer center biorepositories. Furthermore, with sufficient resources, coordination and planning, a population-based virtual biorepository could be created based on linkages between statewide cancer registries, cancer center biorepositories, and community hospitals.

Record linkages similar to the concept tested in this study have been performed between biobanks established for prospective cohort studies and cancer registries in Nordic countries that provide information on cancer incidence and long-term survival for the study participants.⁸ In the United States, the SEER Residual Tissue Repository (RTR) program maintains biospecimens (mostly paraffin-embedded tissue) obtained from three population-based cancer registries in Hawaii, Iowa, and Los Angeles.⁹ The RTR sites collect and catalog specimens that would otherwise be discarded and link them to the cancer registry data. However, programs like the RTR are expensive to maintain, further emphasizing the need for “virtual” biorepositories that link biospecimens stored in existing institutional biorepositories and hospital pathology laboratories with population-based cancer registry data.

The number of records that matched between the two databases varied by year, and was highest for the years 2007 and 2008. The lower proportion of matching records for the years 2005 and 2006 is likely due to the more limited annotation of clinical data for the specimens collected by the UCD CCB during its implementation phase. For 2009, the lower proportion of matching records is likely due to incomplete data in the CCR database. Although medical facilities are required to report new cancer cases to the CCR within 6 months of their diagnosis, the CCR does not usually receive complete data for all of the expected cancer cases until 18–20 months after the end of a given calendar year. This time lag will likely decrease in the future through direct reporting to cancer registries from electronic health records via automated cancer case abstraction or web-based, direct data entry by tumor registrars into cancer registry databases.

Through this pilot linkage, several data elements critical for a successful match between biorepository and cancer registry records were identified. Patient first and last name, date of birth, medical record number, tumor site, and pathology report number were all vital for successful record matches. Both cancer registries and biorepositories should ensure that all data elements are correctly recorded in their databases to facilitate patient identification and biospecimen retrieval.

Data elements for which there was poor agreement between the biorepository and cancer registry databases were also identified. Agreement was poor for date of birth, race/ethnicity, pathology report number, and tumor behavior. For date of birth and pathology report number, lack of agreement was likely due to transcription errors from the medical record into each of the databases. Date of birth may be incomplete (as in month and year only), and is often not included in a pathology report. Pathology report numbers have embedded logic that varies between institutions to describe the year, date, and specimen number. Such variation can make it difficult for tumor registrars to record the pathology report number in the cancer case abstract if they are not aware of the specific institutional conventions. Cancer registries and biorepositories should share the standard

coding systems that are used to record data items in their respective databases and work to incorporate these standards to improve the ability to link specimens.

Records from the UCD CCB were more likely than those from the CCR to be missing race/ethnicity, which is not surprising since this information is not typically included in pathology reports. This discrepancy could be resolved by supplementing UCD CCB records with hospital medical record information. As of June 2010, the UCD CCB gained access to additional patient information through the UCD cancer registry database to enhance the biospecimen clinical annotations.

Tumor behavior codes in cancer registries follow the International Classification for Diseases in Oncology (ICD-O) coding system.⁷ In the UCD CCB database, tumor behavior is coded as pathological status, derived from pathology reports or the medical record, and verified by independent pathologist review. For this pilot linkage, we re-coded the pathological status data in the UCD CCB to align with the ICD-O tumor behavior codes in the CCR database. Differences in how data were abstracted and coded for the two databases made it difficult to align this information. Agreement between these data items could be increased through the internal linkage of cancer center biorepository databases with their institutions' cancer registries.

Surgical “remainder” tissue from patients who donate specimens to the UCD CCB often include both malignant and normal tissue, which is used as control specimens for research projects. The normal tissue specimens were not removed from the linkage process because we wanted to determine the issues that might arise when attempting to link biorepository and cancer registry data. When the benign specimens were excluded from the results, the proportion of records with a match was 79.4%, almost as high as the overall match. The consistency of matches for both normal and benign specimens demonstrates that the ability to identify matches successfully through a probabilistic linkage does not require complete data in both databases.

Still to be explored is the mechanism by which specimens can be provided for research. A survey of pathology laboratory directors in Iowa revealed that most were willing to loan slides and blocks for research as long as confidentiality was maintained, materials were handled and returned properly, and compensation was provided.¹⁰ This is relevant because many patients undergo their initial surgeries in community hospitals, and so true population-based biorepositories must include specimens from these facilities in order to provide a representative research sample. Without access to specimens from community-based pathology laboratories, the sample of biospecimens available from cancer centers will likely be biased towards certain tumor types. For example, the UCD CCB contains a disproportionately high number of prostate cancer specimens because this has been an institutional research focus. Research based on population-based biorepository data should always include comparison with cancer registry data from patients in order to determine what differences may exist between these groups whether the biospecimens are available or not. To ensure the success of virtual biorepositories, cancer centers, cancer registries and community hospitals must collaborate to develop policies on biospecimen sharing, documentation of appropriate informed consent, and protection of donor identity when a biospecimen is released for research. In

addition, public and patient education about the types of research that might be conducted through such linkages and individual rights related to the use of specimens for research should be undertaken in order to ensure that there is public support for these initiatives.

This pilot project was part of an effort to assess cancer center biorepository practices in California and to evaluate support for the creation of a virtual biorepository. The cancer centers generally supported this concept, and contributed time and information to the planning process. Proposed next steps include evaluating the generalizability of research from a virtual biorepository that links a cancer center and a cancer registry; repeating the linkage process with another cancer center biorepository to establish reliability and reproducibility; conducting a sensitivity analysis to determine the number of false-positive and false-negative matches; and demonstrating validity by linking the CCR data with at least two existing biorepositories to procure biospecimens and cancer registry data to answer a molecular epidemiology question. The potential for enhanced information about cancer risk factors and treatment also exists through linking cases to information in electronic health records. The combination of enhanced clinical data from electronic health records with the long-term follow up and survival data available from cancer registries would further increase the value of virtual biorepositories.

In conclusion, linkages between cancer center biorepositories and population-based cancer registries can provide a foundation for virtual biorepository networks and facilitate population-based biospecimen research. The identification of biomarkers that can help tailor cancer treatment regimens by selecting patients who are most likely to receive benefit from a specific therapy is crucial to the advancement of personalized medicine, which could ultimately improve patient outcomes and reduce health care costs.

Acknowledgments

The authors would like to thank Moon Chen, Irmi Feldman, and Kurt Snipes for their contributions to this study.

Author Disclosure Statement

The authors reported no conflicts of interest or financial disclosures.

This study was supported in part by CA114640 and CA153499 but the views represent those of the authors and not necessarily those of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Pro-

gram under contract N01-PC-35136 awarded to the Northern California Cancer Center; contract N01-PC-35139 awarded to the University of Southern California; contract N01-PC-54404 awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement 1U58DP00807-01 awarded to the Public Health Institute.

The ideas and opinions expressed herein are those of the author(s) and endorsement by the State of California, Department of Public Health, the National Cancer Institute, and/or the Centers for Disease Control and Prevention or their contractors and subcontractors is not intended nor should be inferred.

References

1. Sherman ME, Howatt W, Blows FM, et al. Molecular pathology in epidemiologic studies: A primer on key considerations. *Cancer Epidemiol Biomarkers Prev* 2010;19:966–972.
2. Boffetta P. Biomarkers in cancer epidemiology: An integrative approach. *Carcinogenesis* 2010;31:121–126.
3. Goodman MT, Hernandez BY, Hewitt S, et al. Tissues from population-based cancer registries: A novel approach to increasing research potential. *Human Pathol* 2005;36:812–820.
4. Anderson WF. Cancer Surveillance Research. *Cancer Epidemiol Biomarkers Prev* 2009;18:1669–1671.
5. Langseth H, Luostarinen T, Bray F, Dillner J. Ensuring quality in studies linking cancer registries and biobanks. *Acta Oncol* 2010; 49:368–377.
6. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, Becich MJ. A multidisciplinary approach to honest broker services for tissue banks and clinical data: A pragmatic and practical model. *Cancer* 2008;113:1705–1715.
7. Fritz A, Percy C, Jack A, et al. editors. *International Classification of Diseases for Oncology*, 3rd ed. World Health Organization. 2000.
8. Pukkala E, Andersen A, Berglund G, et al. Nordic biological specimen banks as basis for studies of cancer causes and control—More than 2 million sample donors, 25 million person years and 100,000 prospective cancers. *Acta Oncol* 2007;46: 286–307.
9. SEER Residual Tissue Repository Program. [Internet]. Bethesda (MD): National Cancer Institute; 2011 [cited 2011 Sept 21]. Available from: <http://seer.cancer.gov/biospecimen/>.
10. Lynch CF, Kutza JA, Kelley SL, et al. Procurement of population based cancer tissue in Iowa. *Mod Pathol* 1999;12:422–426.

Address correspondence to:

Dr. Rosemary D. Cress

Public Health Institute

Cancer Registry of Greater California

1825 Bell St., Suite 102

Sacramento, CA 95825

E-mail: rcress@crgc-cancer.org