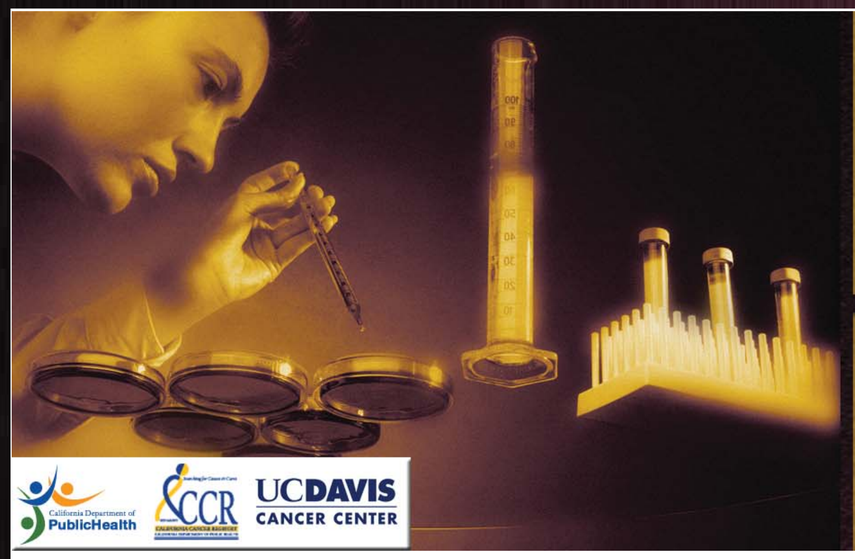

Linking Central Cancer Registries and Institutional Biorepositories to Facilitate Biospecimen-Based Research – A Pilot Study

Margaret E. McCusker, M.D., M.S.,[1] Mark Allen, M.S.,[2] Irmi Feldman,[3]
Allyn Fernandez-Ami, M.P.H.,[2] Kurt P. Snipes, M.S., Ph.D.,[1] Moon Chen, Ph.D., M.P.H.,[3]
Rosemary Cress, Dr.PH.,[2] Regina Gandour-Edwards, M.D.,[3]

1. California Department of Public Health 2. Public Health Institute, California Cancer Registry
3. University of California, Davis Cancer Center

Background:

Central cancer registries have the potential to support population-based biospecimen research by linking cancer surveillance data to existing biospecimens. Cancer registries provide high-quality, population-based data about persons diagnosed with cancer, including their demographic profile, cancer type, first course of treatment and long-term follow up. When these data are linked to biospecimens, population-based studies can be conducted to evaluate the molecular profiles of tumors; describe the molecular epidemiology of newly-identified oncogenes and their impact on recurrence and survival; study the molecular epidemiology of rare tumors and tumors among specific population subgroups, including those most affected by health disparities, and validate these findings by comparing data on patients with and without biospecimens.

Purpose:

To determine if existing biospecimen records from the University of California, Davis Cancer Center Biorepository (UCD CCB) could be reliably linked with patient records from the California Cancer Registry (CCR). This project was a pilot study designed to test the feasibility of linking biorepository databases with the CCR database and was part of a larger project to develop plans for a biospecimen research network in California.

Methods:

We performed a probabilistic data linkage between 3,092 UCD CCB biospecimen records and 3.3 million CCR records based on standard CCR data linkage procedures. UCD CCB records for the years 2005-2009 and all cancer cases reported to CCR through 2009 were included in the linkage. Table 1 lists the variables from each database that were included in the linkage. Only UCD CCB records with a unique value for medical record number, tissue site, and pathology specimen date were used since most individuals who donated biospecimens had more than one specimen in the biorepository. UCD CCB race/ethnicity, tissue site and tumor behavior variables were re-coded to align with CCR codes. The linkage process comprised six sequential comparisons of the two data sets, which accounted for possible differences in how variables were recorded, such as typographical errors or variations in coding from the medical record that were not true differences. Variables with the same value in the UCD CCB and the CCR databases received a positive agreement weight, and those that were different received a negative weight. The weights of all of the variables were added, and those with high total weights were considered matches. If a patient had two specimens from two separate occasions in the UCD CCB database, both specimens would be counted as matches.

Table 1: Variables from Each Database Included in the Linkage

Variable	UC Davis Cancer Center Biorepository	California Cancer Registry
First Name	X	X
Middle Initial	X	X
Last Name	X	X
Maiden Name		X
Gender	X	X
Date of Birth	X	X
Race/Ethnicity	X	X
Medical Record Number	X	X
Tissue Site	X	X
Tumor Behavior	X	X
Pathology Specimen Date	X	
Pathology Report Number	X	X
Date of Diagnosis		X

Prepared by the California Cancer Registry, California Department of Public Health, Cancer Surveillance Section.

Results:

For the years 2005-2009, 1,040 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 (Table 2). For the major variables used to link records between the databases, 99.4% of matched cases had the same value for gender, while only 42.8% had the same value for tumor behavior (Table 3). Table 4 shows the number of records in the linkage which were identified in both databases by cancer site. Overall, record matches were highest for cancers of the cervix (100%) and testis/other male genital system (100%). Matches were lowest for cancers of the skin (20%) and bones/joints (33.3%). For the most common cancers, matches were highest for lung and respiratory system (93%), breast (91.7%), and colon and rectum (89.5%) and lower for prostate cancers (72.9%).

Table 2: Matches by Year Between the UCD CCB and CCR Databases

Year	Used	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 (2005-2009)	1,040	844	81.2%

Prepared by the California Cancer Registry, California Department of Public Health, Cancer Surveillance Section.

Table 3: Agreement Between Variables in the UCD CCB and CCR Databases, 2005-2009 (n=844)

Variable	Number in Agreement	% Agreement
Gender	839	99.4%
Last Name	927	98.0%
First Name	812	96.2%
Tumor Site	724	85.8%
Medical Record Number	707	83.8%
Ethnicity	689	81.6%
Date of Birth	546	64.7%
Race	545	64.6%
Pathology Report Number	478	56.6%
Tumor Behavior	361	42.8%

Prepared by the California Cancer Registry, California Department of Public Health, Cancer Surveillance Section.

Table 4: Records Matching by Cancer Site in the UCD CCB and CCR Databases, 2005-2009 (n=1,040)

Description	# Cases Used	# 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 Including Heart	48	34	70.8%
Miscellaneous	18	12	66.7%
Bones and Joints	6	2	33.3%
Skin	5	1	20.0%

Prepared by the California Cancer Registry, California Department of Public Health, Cancer Surveillance Section.

Conclusions:

The test linkage between the UCD CCB and CCR databases demonstrated that existing biorepository data can be successfully linked with cancer registry data to identify biospecimens for population-based biospecimen research. Critical variables for such linkages include first and last name, date of birth, facility medical record number, cancer site, and pathology report number. Based on the results of this pilot study, improvements in data quality and completeness for these variables within both the UCD CCB and CCR databases will help to improve the success of future linkages. In addition, a review of how the data are coded in each database would help to determine if standardized coding for variables across both databases could improve the proportion of matched cases. Linkages between existing biorepositories and cancer registries can foster productive collaborations between these entities, and provide a foundation for virtual biorepository networks to support population-based biospecimen research.

This work was funded in part by CA U01CA114640 (AANCART) but the content is solely based on the presenters/authors and does not necessarily reflect the views of the National Cancer Institute.