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 biospecimen 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.

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 (28%) 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%).

Conclusions:
The test linkage between the UCD CCB and CCR databases demonstrated that existing biospecimen 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 biospecpositories and cancer registries can foster productive collaborations between these entities, and provide a foundation for virtual biospecitory networks to support population-based biospecimen research.