Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

_____ _____

Katherine Ann John Date

USING CAPTURE-RECAPTURE METHODOLOGY TO ESTIMATE ADOLESCENT AND ADULT CONGENITAL HEART DEFECT (CHD) PREVALENCE IN FIVE METROPOLITAN GEORGIA COUNTIES: 2008-2010

BY

Katherine Ann John

Master of Public Health

Epidemiology

_____ [Chair's Signature]

Carol Hogue, PhD, MPH

Committee Chair

_____ [Member's Signature]

Cheryl Raskind-Hood, MPH, MS

Committee Member

_____ [Member's Signature]

Wendy Book, MD

Committee Member

USING CAPTURE-RECAPTURE METHODOLOGY TO ESTIMATE ADOLESCENT AND ADULT CONGENITAL HEART DEFECT (CHD) PREVALENCE IN FIVE METROPOLITAN GEORGIA COUNTIES: 2008-2010

By

Katherine Ann John

Bachelor of Science

Bucknell University

2011

Thesis Committee Chair: Carol Hogue, PhD, MPH

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Epidemiology

2015

Abstract

USING CAPTURE-RECAPTURE METHODOLOGY TO ESTIMATE

ADOLESCENT AND ADULT CONGENITAL HEART DEFECT (CHD)

PREVALENCE IN FIVE METROPOLITAN GEORGIA COUNTIES: 2008-2010

By Katherine Ann John

Purpose: To determine the congenital heart defect (CHD) prevalence in five metropolitan counties (Clayton, Cobb, DeKalb, Fulton, and Gwinnett) in Georgia from 2008-2010 using capture-recapture methodology.

Method: Using data from Children's Hospital of Atlanta (CHOA), Sibley Heart Center Cardiology, Pediatric Cardiology Services (PCS), Grady Health, Emory Healthcare including St. Joseph's Hospital, and Georgia Medicaid claims, capture-recapture (CR) methodology and logistic regression were employed to estimate the prevalence of CHD for both adolescents, aged 11-20, and adults aged 21-64, in five metropolitan Atlanta, Georgia counties. From this, the number of CHD cases that were missed were estimated by these data sources from January 1, 2008 to December 31, 2010.

*Results***:** Altogether 1,858 adolescent cases were captured from at least one "adolescent" database (CHOA, Sibley, PCS, and Medicaid), and 3,183 adult cases were captured from at least one "adult" database (Emory Healthcare, St. Joseph's Hospital, Grady Health, and Medicaid). The estimated number of adolescents (aged 11-20 years) with CHD and living in the 5 metropolitan Atlanta counties in Georgia was 3,718 (95%CI: 3,471 - 4,004) for a prevalence estimate of 7.85 per 1,000 population aged 10-19 in 2010. The number of adults with CHD aged 21-64 years was estimated to be 12,969 (95%CI: 13,873 -18,915) for a prevalence estimate of 6.08 per 1,000 population aged 20-64 in 2010.

Conclusion: Despite the need for lifelong care, adults with CHDs are being lost within the healthcare system. Public health initiatives should focus on the high proportion of adult CHDs retained in adolescent care. Lack of referrals and patient retention in adolescent care provides context for the need of more specialized adult congenital care units and to mandate policies, such as patient referrals, to assist physicians with coordinated transfer of patients to adult care.

USING CAPTURE-RECAPTURE METHODOLOGY TO ESTIMATE ADOLESCENT AND ADULT CONGENITAL HEART DEFECT (CHD) PREVALENCE IN FIVE METROPOLITAN GEORGIA COUNTIES: 2008-2010

By

Katherine Ann John

Bachelor of Science

Bucknell University

2011

Thesis Committee Chair: Carol Hogue, PhD, MPH

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master in Public Health

in Epidemiology

2015

Acknowledgements

Thanks to Dr. Carol Hogue and Cheryl Raskind-Hood for advising and mentoring me through the thesis process. Thanks to Drs. Wendy Book and Bill Mahle for their guidance and support. I would also like to thank Jill Glidewell and Pam Costa with the Centers of Disease Control and Prevention and the members of the Analytic Work Group of this pilot Congenital Heart Defect Surveillance Project from the State Health Departments in Massachusetts and New York.

Table of Contents

List of Abbreviations:

ACC	**American College of Cardiology**
ACHD	**Adult Congenital Heart Disease**
AIC	**Akaike Information Criteria**
ALS	**Amyotrophic Lateral Sclerosis**
ARP	**Alcohol Related Problems**
CDC	**Centers for Disease Control and Prevention**
CHD	**Congenital Heart Defects**
CHOA	**Children's Healthcare of Atlanta**
CMS	**Centers for Medicare and Medicaid Services (Medicaid claims data)**
CR	**Capture-Recapture**
DF (df)	**Degrees of Freedom**
HCUP	**Healthcare Utilization and Cost Project**
IDDM	**Insulin-Dependent Diabetes Mellitus**
MACDP	**Metropolitan Atlanta Congenital Defects Program**
MAX	**Medicaid Analytic eXtract files**
MCO	**Managed Care Organization**
MSIS	**Medicaid Statistical Information System**
NCDDD	**National Center on Birth Defects and Developmental Disabilities**
NIDDM	**Non-Insulin-Dependent Diabetes Mellitus**
PATCH	**Provider Action towards Congenital Heart Defects**
PCS	**Pediatric Cardiology Services**
PPACA	**Patient Protection and Affordable Care Act**
ResDAC	**Research Data Assistance Center**

SLE	**Systemic Lupus Erythematosus**
SSA	**Social Security Administration**
STS	**Society of Thoracic Surgeons**
VAERS	**Vaccine Adverse Event Reporting System**
VBA	**Veterans Benefits Administration**
VHA	**Veterans Health Administration**

CHAPTER I: Background & Literature Review

Background

Congenital Heart Defects (CHD) are the most common type of birth defect, affecting roughly 1% of births per year.[1] CHDs are problems with structure or function of the heart that are present at birth and can involve all parts of the heart, including the interior walls of the heart, the valves inside the heart, or the arteries and veins that carry blood to the heart from the body.[2] In administrative data, CHDs are coded separately for each abnormality of structure or blood flow, so that an individual can have multiple codes to describe one CHD.[3] These heart defects can range from simple defects with no symptoms to life threatening conditions. Due to the variability in CHD classification, both the prevalence and incidence of CHDs have been difficult to assess. Inaccurate estimates of prevalence and incidence affect estimates of morbidity, mortality, and health care costs attributable to CHDs.

To reduce CHD nomenclature and increase correct classification, physicians from the Society of Thoracic Surgeons (STS) and the European Association for Cardio-Thoracic Surgery created the International Pediatric Congenital Cardiac Code in 2000.[4] The STS coding system is widely used due to its simple, specific, standard nomenclature. It contains 2- to 4-digits, and partially aligns with the ICD-9CM.[3] This clarification streamlines both the diagnosis and management for individuals with CHD. It is estimated that greater than 85% of patients diagnosed with a CHD survive into adulthood owing to significant advances in diagnosis and management of CHDs.[5,6] With these advances comes the need for data on later outcomes in this population, as well as a clearer estimate on the enumeration of adults living with CHD.

The Healthcare Utilization and Cost Project (HCUP), the largest longitudinal data collection project, evaluates U.S. hospital utilization and costs. Since inpatient hospitalization is often necessary to treat people with CHD, the HCUP Nationwide Inpatient Sample (NIS) database was used as a data source to derive population estimates for the year 2004. HCUP is based on community hospitals, excluding long-term hospitalizations from calculations and utilizes hospital discharges as the unit of analysis. In the U.S. in 2004, 46,500 of the 139,000 birth defect hospitalizations were due to circulatory or cardiovascular anomalies.[7] Of these cardiovascular hospitalizations, 34% of them resulted in over half of the health care costs attributed to the total birth defect hospitalizations.[8] These costs need to be taken into consideration as the demographics of this population change.

In 2000, a CHD prevalence study was conducted in Quebec, Canada to determine population estimates of CHD across the life-span.[8] Quebec is unique in that it enjoys universal healthcare coverage with each individual assigned a unique Medicare ID number at birth which tracks diagnoses and health services accessed over the individual's lifetime. The Quebec Congenital Heart Disease Database was created by merging and cleaning data from administrative databases, hospital discharge summary databases, and the Quebec Health Insurance Board and Death Registry Data. The final Quebec CHD database encompasses 28 years of longitudinal data on all individuals with CHDs over the time period of 1983-2000.[8] The goal of this study was to use this longitudinal database to both estimate lifetime prevalence of CHD, while also comparing the number of adults with CHD to the number of children with CHD in the Quebec population from 2000 to 2010. Using the unique patient identifiers, data updates were requested for the

same administrative sources as used in the 1983-2000 study up to the year 2010. The unique patient ID's were also used to de-identify the data minimizing the number of duplicate records and to capture individual subjects for correct encounter linkage for prevalence estimates. Using these data, the prevalence of CHD in children in 2010 was 13.11 per 1,000, while the prevalence was 6.12 per 1,000 in adults. From 2000 to 2010, CHD prevalence in children increased by 11%, and by 57% in adults with adults representing two-thirds of the CHD population.[8] Improved care, decreased mortality and/or improved diagnosis over the life-span are likely contributors to increasing prevalence of CHD in adults and children found in this study.[8] Increased prevalence of the aging population presents the possibility that adults with CHDs may also have comorbidity, adding to the disease burden of this population.[8]

While there is robust evidence of CHDs detected at birth, there remains no population-based surveillance data on prevalence beyond early childhood in the United States.[1] Estimates of prevalence in adulthood would provide a clearer picture of the disease burden on U.S. health care utilization and attributable costs, morbidity, mortality, and non-health care costs. Extrapolation from Canadian data to the U.S. population has provided estimates that suggest roughly 2 million people, including both adults (~959,000-1.5 million) and children (~975,000-1.4 million), are living with CHDs.[9] Population-based surveillance of CHDs would provide data on the magnitude of the condition, distribution of the condition geographically, natural history of the condition, and changes in prevalence over time based on the evolving population and prevention strategies/activities.[1]

Access to Care

Since there is no population-based surveillance of CHD across the lifespan in the United States like there is in Canada, we must rely on prevalence estimates from the year 2000 to determine the number of adults and children living with CHD in the US.[9] The current estimate is that roughly 1 million US patients are living with CHD (adults 800,000 and children 600,000); most of them require lifelong care with over half of this population requiring specialist treatment according to physician guidelines for management and care.[5,9] Gaps in care are a public health burden affected by ongoing racial/ethnic disparities, economic disparities, other social disparities, and geographic differences. Gaps in care are a detriment for many health conditions, and there is concern that gaps in care for the CHD population will result in large inequalities between groups as they age. A study conducted by the Adult Congenital Heart Association (ACHA) found that 42% of study participants, which included a large proportion of highly educated adults from different CHD care programs in the U.S., had a greater than three-year gap in care.[5] Adult CHD patients often have interruptions in care, multiple gaps, where the first gap of care is most commonly recognized during the late teen years, during the transition from pediatric to adult-oriented medical care.[5] The mean age of CHD patients within the first gap was 19.9 years of age, with the most common reported reasons for this gap in cardiology care being: "felt well," "did not need follow-up," "not receiving medical care," "moved," or "changed or lost insurance."[5] There is little information available about the underlying reasons for these gaps or about individuals living with CHD who are being missed in the healthcare system.

Access to appropriate care and elucidating the key barriers to healthcare access including unemployment, lack of adequate health insurance, transitioning of care from childhood to adulthood, and lack of proximity to a specialized care center are essential in ensuring optimal health services for patients living with a CHD.[1] While access to care is a significant barrier to estimating the true number of CHD patients, the longevity of the U.S. population presents another significant challenge. Living longer with the disease creates ample time for patients to become lost to follow-up or discontinue recommended care from childhood to adulthood. In a study looking at U.S. inpatient hospitalizations for congenital heart defect admissions from 1998-2010, admission counts by age as well as other characteristics showed that adults made up only 36.5% of the hospitalizations captured during the latter era of the study.[10] The majority of both children and adults with CHD had either public insurance or were not insured, and the mean length of a hospitalization stay was greater for children compared to adults in the latter era, 17 days compared to 5.8 days, respectively.[10] While admission counts for both children and adults increased when comparing hospitalizations from 1998-2004 to 2004-2008, simple defects make up a greater percentage of adult admissions compared to children.[10] The frequency of hospitalizations for adults with CHD is likely due to better procedures, the aging population, and the accumulating comorbidities found in this group.[10] These data reflect hospitalization level data rather than patient-level data and further reflect the burden of CHD admissions as opposed to count estimates of patients with CHDs.[10] Further research is needed as to the effect of the adult CHD population on resource utilization and healthcare delivery.

With the passing of the 2010 Patient Protection and Affordable Care Act (PPACA), it is likely that there will be a shift towards improved access and utilization of health care insurance and services. Of the $149,871,595 which was awarded to Georgia for primary care services, extension of operation hours, hiring of additional providers, and renovating or building new clinical spaces, $5,176,702 was awarded to Georgia health centers to help enroll uninsured Americans in the Health Insurance Marketplace.[11] Under this new healthcare law, children can now be maintained on their parent's health insurance policy until they turn 26 years old. Thanks to this provision, 123,000 young adults in Georgia who would otherwise have been uninsured have gained coverage nationwide.[11] The PPACA also no longer allows insurance companies to deny coverage to individuals based on pre-existing conditions. In 2013, of the 4,323,897 non-elderly Georgia citizens with a pre-existing condition, 613,253 are children[11] and although Georgia has not expanded Medicaid since the Health Marketplace's first open enrollment period in October 2013, the PPACA will help increase insurance coverage for those Georgians and could possibly lead to more accurate estimates of disease prevalence's.

Capture-Recapture Methodology

The capture-recapture (CR) method attempts to generate estimates to account for incomplete ascertainment of cases overlapping from one or more distinct sources.[12] These methods were initially developed in ecology to estimate the size of wildlife populations and have been readily applied to epidemiological studies for estimates of true sample size.[13] The most basic theory is that each source of data is a simple random sample of the total study population; for instance, the appearance of an individual case's name on a list

does not influence the name appearing on any other list, meaning that each source is independent.[14]

The term "source" is widely used in applications of CR methods, especially in epidemiological studies denoting a list of cases; however, these lists are not frequently standardized on case ascertainment.[12] Lists of cases can be utilized from death records, birth certificates, disease registries, laboratory reports, medical billing or clinical records and educational data, all of which may apply different methods to identify, obtain, and report information on cases. A clear, precise, and accurate case definition is required to determine the number of unique captures across data sources. In addition, disease definitions across sources must comprehensively encompass the spectrum of disease manifestation to maintain a consistent probability of being captured; all data sources, characteristics pertaining to those sources, and patterns of interest should be explicitly stated at the beginning of the CR method.[15] Moreover, critical decisions regarding how to define analytic sources and determine if certain sources should be pooled should be made before any data are ascertained. Once the sources are identified, relationships between each pair of sources (i.e., positive and negative dependence) must be addressed prior to modeling.[15] Reporting dependency between sources may assist in deducing the direction of bias in implausible estimates or may support that the derived estimates were plausible when exploring prevalence or incidence of disease.[12]

The simplest CR model is the two-sample model. Within a prescribed catchment area, two samples are obtained and a unique identifier is used which de-duplicates entries between the samples yielding the number of unique individuals in the catchment area. Once the sources are de-duplicated, the number of unique individuals captured in source

1, source 2, and in both sources is determined. Using the number of individuals captured in each of the two sources and the number of individual cases captured in both sources, the number of individuals not captured in either source can be estimated. For this estimate to be accurate, several assumptions must be made: (1) there is no change to the population during the capture time period (closed population); (2) there is no loss of individuals; (3) each individual has the same probability of being captured; and (4) the samples are independent of one another.[13] Under the assumption that the two sources are independent, the number of cases missing between the two sources can be estimated by multiplying the number of cases found in source 1 by the number of cases found in source 2 and dividing by the number of individuals captured by both sources. Once the number of missing cases is determined, the total number of individuals can be estimated through the addition of the individuals from both sources, from each individual source, and the estimated missing cases.[12]

When data from multiple sources are available, the analysis becomes more complicated. Hook urges investigators to present results for all two-source estimates including 95% confidence intervals (CI).[15] The CIs are derived by considering each source individually to all other pooled sources, as well as, two-source estimates of each source against the others.[15] Fienberg was the first to develop an approach for multiple sources by modeling a multiple source CR approach through an incomplete 2^k contingency table with one unobservable cell.[16] Using this approach, log linear modeling is the most common modeling strategy employed and is used to handle dependency between the sources. There are four approaches investigators can take when using log-linear modeling strategies: (1) log-linear analyses in substrata defined by combination of

suspected covariates like age, gender, etc.; (2) a single log-linear analysis by adjusting simultaneously for defined levels of pertinent covariates; (3) log-linear analysis of the entire population assuming covariates affect data structure of the entire population; and (4) Bayesian approaches.[15] Interaction terms are used to model local dependence between sources, and there is a natural assumption that there is no k source interaction term for the multiple models.[17] The four general types of models to incorporate these dependencies include: 1. *independent model* which assumes all sources are independent from one another; 2. *models equivalent to the two-way CR*; 3. *models that assume all possible interactions;* and 4. *the saturated model*.[12] How well the various log linear models fit to the observed cells is assessed using the deviance statistic, Akaikie Information Criterion, Bayesian Information Criterion, and Goodness of Fit statistics.[17-19] Akaikie Information Criterion is the log likelihood of the model evaluated at the vector of the unknown parameters, and the number of distinct parameters ($AIC = -2\log(L((\beta)) + 2w$). A model producing a small AIC has a better fit and is more parsimonious.[16]

Utility of Capture-Recapture (CR) Method in Public Health

One of the first applications to human populations was conducted in 1949 for the estimation of birth and death rates in Calcutta, India.[20] Sekar and Deming used a birth/death registry alongside a house-to-house canvasing list to estimate the number of births and deaths occurring in1945 and then again in 1946. While the ascertainment of these lists have been modernized since the 1940s, registrar and house canvassing information were the only available sources for investigators to use at the time, but the statistical/theoretical applications employed are still very much applicable today. Sekar and Deming warn that CR estimation using two sources oversimplifies the situation and

that there are inherent weaknesses in case ascertainment that should be acknowledged, and if possible, corrected for in analyses.[20] The authors provide the following possible reasons for incomplete investigations: unclassified entries due to illegibility, incompleteness of entry, or failure of the investigator to properly ascertain information, individual's movement either permanently or temporarily to a new residence, non-residents having events in capture institutions and precision of sampling methods.[20] It should be recognized that the occurrence of these events will all invariably affect the precision of the estimates using CR techniques.

More recent CR applications include the estimation to the size of undercount in censuses, estimating the number of duplicate records on a list or database, refinement of prevalence or incidence estimates derived from attempted exhaustive population surveys, attempted evaluation of source completeness, and attempts at deriving plausible upper and lower limits on the total affected.[15,21] One notable example is the use of CR to assess the reporting completeness of a passive reporting system, along with the possible assessment of risk of event. The Vaccine Adverse Event Reporting System (VAERS) is a passive surveillance reporting system used to monitor vaccine safety and unfortunately is limited in completeness by severity of event, proximity in time of the event to vaccination, and preexisting awareness of the event to vaccination.[22] VAERS, a retrospective cohort study, and a case-control study were used as three independent sources in a CR analysis. Risk estimation was conducted using the total estimated cases divided by total person-time for each pre-specified time interval. The applicability of use of these sources was limited by the following observations from the researchers: a possible dependency in managed care organization (MCO) reporting that could not be

evaluated due to lack of information in the VAERS database, severity of disease could have affected ascertainment rates, public awareness of the correlation between vaccination and intussusception (prolapse of a section of bowel resulting in bowel obstruction observed in children vaccinated for rotavirus) could have affected vaccination rates, concise disease definition, and possible confounding from unavailable variables (type of healthcare, insurance, etc.).[22] The applicability of using these sources for a different disease seems unlikely and is further complicated by the use of VAERS as a source in the analysis.

Any approaches to prevalence estimation rely on complete ascertainment of true cases, and the investigator must fully investigate the data's ascertainment history and structure especially when using sources of convenience.[15] Hook and Regal analyze McGilchrist and colleagues' aim at using CR to estimate the number of measles cases occurring in children under the age of 10 in the Blacktown area of Sydney, Australia during the period of June 1 to December 30, 1993.[23] McGilchrist's choice in using four sources: diagnoses from doctors, reports from hospitals, laboratory specimens, and "other mechanisms" lends concern to the accuracy of diagnoses, relationships/dependencies between sources, and possible targeting to a source population that is separate from the other data sources.[15] The anomalous nature of source "other" creates numerous issues with estimation. "Other" cases found may not be true cases and could contribute to matching problems because they may either target a source population separate from the other sources (geographical, socio-economical) implying variable catchability or could include individuals known not to be reported by a register for not meeting the case definition.[15] Investigators should be aware of possible anomalies in the data sources, and

the use of sources of convenience may cause large extrapolations to estimations. Discarding the "other" source and using three sources for CR could provide clarity and useful prevalence or incidence estimations.

Epidemiology revolves around evaluation of efficiency and effectiveness in relation to surveillance systems and it is difficult, if not impossible, to improve on efficiency of a system without first knowing its completeness.[24] CR methods adjust multiple population estimates identified through multiple incomplete sources to reflect census undercount or ascertainment level of the monitoring system. An adolescent injury monitoring system was evaluated using the following four sources: (1) a 1-month student recall; (2) a 4-month student recall; (3) medical excuses; and (4) attendance records. Through two source CR analyses, it was determined that the 1-month and 4-month student recalls were extremely similar and did not capture a proportionate number that the other did not, prompting the authors' decision to pool the sources. Log-linear modeling supported this decision in that the best fitting model was the one controlling for interaction between the 1-month and 4-month sources.[24] To determine the most efficient source combination, investigators are often faced with a choice of combining sources to attain a higher degree of precision.[24] Trade-offs exist and the cost of case-finding should be balanced with the precision needed to evaluate a particular disease of interest.

While compulsory reporting to registries is the broadest method used to obtain estimates of incidence and prevalence of a population's disease burden, it is often expensive and does not fully lend itself to complete enumeration of cases.[25] Births and deaths tend to be well documented; however, variables that could be related to the capture (severity of disease or sources used in the reporting) may be missed from a

registry and could lead to biased inferences.[25,26] CR methods offer the potential to reduce both the costs of disease registers and the likelihood of attaining biased estimates of incidence and prevalence of disease.[26] CR techniques offer a less expensive and often a more informative approach.[25]

Studies Utilizing Capture-Recapture (CR) Methods to Estimate Prevalence

Disease registries are collections of information about individuals with specific conditions that provide researchers, clinicians, and health-care professionals with information to better understand certain diseases, track trends, and identify possible treatment measures.[27] The Amyotrophic Lateral Sclerosis (ALS) Act, passed in October 2008, called for the establishment of a national registry of patients with ALS.[28] CR methods were employed on data collected in Georgia from Emory Healthcare, the Veterans Health Administration (VHA), the Veterans Benefits Administration (VBA), Medicare, and Georgia mortality records to estimate the period prevalence of ALS in metropolitan Atlanta from 2001-2005.[29] The data were collapsed into four sources and using a unique identifier, cases were linked across source to determine the number of unique cases found in the each of the sources. Both two-source CR to assess dependency between the sources and log-linear modeling under the Poisson distribution stratified by age were performed to assess the completeness of each source and to estimate the prevalence of ALS.[29] A saturated model, a model containing 4 main sources and 3-way interactions stratified by age and dichotomized race, was chosen as their final model. This model yielded a total case population estimate of 880 (95% CI: 816, 965), a 5-year prevalence estimate of 38.5 per 100,000 (95% CI: 35.66, 42.19) using 2003 census information from metropolitan Atlanta, and an estimate of 273 missing cases from the

original 798 cases used in the CR method.[29] The two-way capture-recapture analysis showed strong positive source dependency and case-source heterogeneity which was addressed in log-linear modeling by stratification and by inclusion of interaction variables in the final model. Although CR methodology is a complex and interactive process, this study provided information on data gaps and helped facilitate the establishment of an effective national ALS registry.[29]

Ascertaining community health problems through screening and questionnaires provides important information, including community costs for the health problem and treatment options for health-care professionals; however, this methodology is not feasible on a national or regional scale.[30] CR methods have been utilized to generate prevalence estimates on a variety of public concerns including Alcohol Related Problems (ARP) in a rural, Italian community using multiple incomplete lists. A strict case definition for individuals limited to those who received treatment in 1997 was used to identify cases found in 4 sources: (1) self-help volunteering groups; (2) psychiatric ambulatory; (3) public alcohol service; and (4) hospital discharge records.[30] Linkage was conducted with a unique identification code and both two-source CR techniques and log-linear models were fit to the data. The goodness of fit statistics for determination of the final model revealed that age was responsible for the heterogeneity in capture found in the two-source capture analysis. The saturated model, containing all 4 sources, 3-way interaction terms, and stratified by age and gender, was chosen as the prediction model which yielded an estimate of 2,500 patients with ARP and a prevalence of 19 per 1,000 individuals older than 15 years.[30] Limitations of this study include the strict case definition employed which might have led to an underestimate of cases in the community through limited

capture sources.

Many individuals with chronic diseases have a longer life expectancy due to better technology and lifestyle changes, making prevalence estimates for these groups a priority to determine trends, monitor complications, and provide essential information to health planners.[31] For example, as an alternative to implementing a diabetes monitoring system, in 1988, Italian researchers applied CR methods to estimate the total number of cases in Casale Monferrato with a diagnosis of insulin-dependent diabetes mellitus (IDDM) or non-insulin-dependent diabetes mellitus (NIDDM) to determine the prevalence rate and the 95% confidence interval surrounding this estimate.[31] Four sources (diabetic clinics/family physicians, computerized database containing prescription records, hospital discharge records, and reimbursement lists for reagent strips/insulin syringes) were used for two-way CR analysis and log-linear modeling. Dependencies between sources found in the two-way CR analysis indicated that the best log-linear model was the one containing all 4 sources, interaction terms, and stratified by pattern of treatment (a possible confounding variable). This model estimated the prevalence of diabetes in Casale Monferrato at this time to be 2,586 cases for an adjusted prevalence rate of 2.77% for residents (95% CI: 2.44, 3.10).[31]

Due to increased globalization and increased access to travel, disease introduction to endemic areas poses a great public health risk. Dengue fever is a common mosquito-borne viral disease that has increased 30-fold in the past 50 years due to increasing geographic expansion and transmission through travel.[32] Surveillance reporting has been established in Europe considering the risk of epidemic for imported dengue cases. In France, CR methodology was employed on its metropolitan 3-source surveillance system

(mandatory reporting by physicians and biologists, laboratory reporting, and enhanced surveillance reporting in vector established departments) to estimate annual incidence during each year from 2007-2010 and a combined estimate over the entire period.[32] Two-source CR revealed that the enhanced surveillance system was highly dependent on the other two sources and was dropped from further analysis.[32] Due to the high dependency, Chao's estimator and stratification by geographic area, year, and time of year was conducted to determine incidence by the laboratory surveillance network and the mandatory reporting network. Using this method, 327 cases were revealed over this 4-year period and of these 234 cases occurred in 2010.[32] Completeness of the mandatory notification network and laboratory network was found to be 10% and 40%, respectively.[32] Use of the CR methodology allowed for estimation of completeness of the two notification systems and provided information for further implementation strategies, monitoring spatial and temporal trends, and assessment of risk by geographic origin.[32]

CR methodology is an excellent strategy for diseases with complex etiology and diagnostic complexity, requiring a multitude of findings from various clinical settings. For example, due to the United States' fragmented health care system and lack of autoimmune disorder surveillance, methods to ascertain incidence and prevalence of diseases like Systemic Lupus Erythematosus (SLE) are in demand.[33] For CR analysis, information on individuals diagnosed with SLE in a Michigan-based study during the three year surveillance period from January 1, 2002 through December 31, 2004 was obtained from 4-sources:hospital data; rheumatologist data; nephrologist/dermatologist data; and End-Stage Renal Data System.[33] Using two-source analysis, an additional 7 cases were determined to be contained within the source population.[33] The overall age-

adjusted prevalence from the four-source model was 72.8 per 100,000 persons (95%CI: 70.8, 74.8).[33] Stratification by age and race found substantially different prevalence estimates, thereby providing opportunities for comparing estimates between the groups and more information for diagnostic methods to accommodate these disparate groups. This CR analysis showed that collaboration between registries is needed as well as infrastructure improvement to capture the highest risk groups.

Limitations for Capture-Recapture (CR) Methodology

Although the CR method originated with wildlife studies, it has been widely applied to human studies. Some main differences between these two populations are that wildlife population estimates include many trappings, have a general time ordering, and the model provides insight on animal behavioral responses. In animal settings, a trap or net is placed in the study area and at first trapping, animals are marked with a unique tag and at subsequent trappings are recorded if recaptured or tagged with a unique tag if unmarked.[17] When applying these methods to human populations, trapping samples are regarded as lists for ascertainment data. Three main differences between wildlife and human captures are that (1) usually wildlife sampling contains a multitude of traps whereas epidemiological studies usually have 2-4 lists available for case ascertainment, (2) generally no time ordering exists or will vary by individual in epidemiological studies where animal experiments have a natural time ordering, (3) and lastly animal studies use identical trapping mechanisms where animal behavioral response to capture can be noted and used in analysis, but human populations must rely on different types of ascertainment sources that are utilized to search for cases.[17]

When applying the CR method with humans, the assumption of loss of individuals is dependent on the quality of the data records and the ability to create a unique identifier for matching individual cases. The assumption of homogeneity within the population is linked with the probability of being captured, which in epidemiological studies is through the use of a case definition. If the case definition is not applied consistently or does not contain the full spectrum of the disease, the probability of being captured would vary over the spectrum of the disease.[34] It is likely that more severe cases captured by one source are more likely to be captured by the other, for example a severe case of cardiovascular problems is more likely to be admitted to a hospital, and then, if this case dies, the death is more likely to be recorded correctly as death due to cardiovascular problems, leading to dependence between the two sources.[26] Members of the population can also differ considerably in probability of case ascertainment by geographic region and socioeconomic variables.[12]

Thus, the assumptions regarding sample independence in CR methods are generally false due to the nature of the population and the scope of our health care system including hospital admissions, doctor records, and patient referrals.[13] Positive dependence between sources provides an underestimation of the total population, while sources that are negatively dependent, such as mutually exclusive databases based on geographic region, typically result in an overestimation of population size.[34] The concept of dependency between sources can be illustrated through the use of a Venn diagram (Figures 1 and 2). The box represents the total population that is being measured and inside are the 2 overlapping sources (Source 1 and Source 2). The degree of overlap, where a large overlap is indicative of positive dependency (Figure 1) and a small overlap

represents negative dependency (Figure 2) portrays the distribution of case classification. The size of the intersection directly affects the area containing the surrounding missing and thus the calculated total population. The inclusion in one source has direct causal effect on his/her inclusion in other sources. Non-independence can be caused by list dependence where dependence is conditional on the individual. Another cause of unequal catchability is based on heterogeneity between individuals, which is a phenomenon sometimes seen in the aggregation of two independent 2X2 tables resulting in a dependent table. These two dependences are difficult to disentangle during data analysis and result in bias, which can lead to both underestimates of samples that are positively dependent and overestimates of negatively dependent samples.[17]

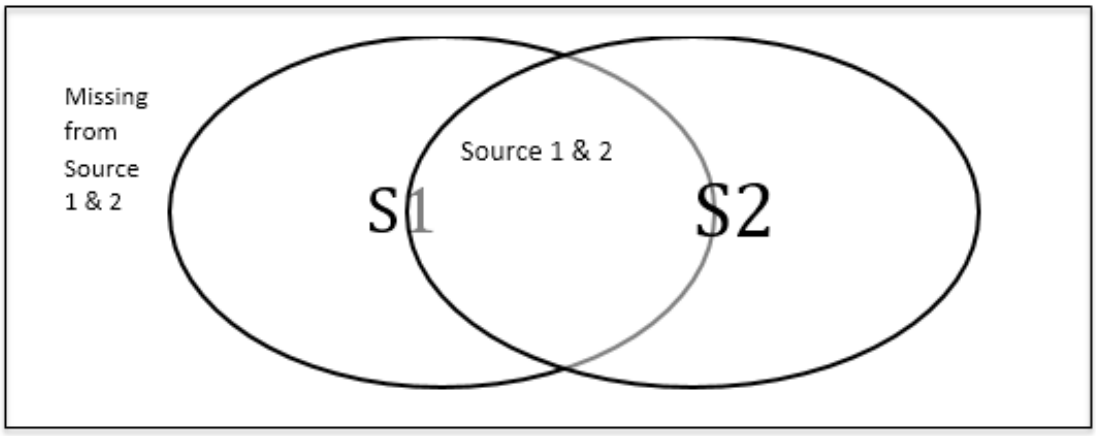

Figure 1. Venn diagram illustrating positive dependence.

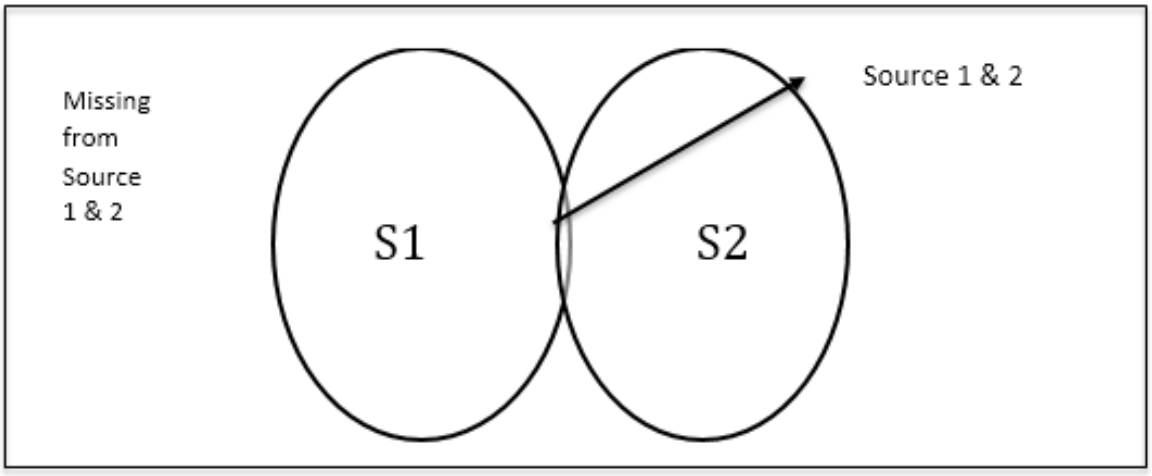

Figure 2. Venn diagram illustrating negative dependence

Log-linear models are being increasingly used by researchers to handle dependencies between sources,[17] for estimates using more than 2 sources, and to handle pertinent covariates.[15] However, due to the ease of incorporating factors into these models, computational burden can arise through the addition of sample periods and strata.[16] Difficulty may arise in the identification of pertinent covariates and even if pre-existing knowledge exists, data may not be readily available to control for these variables.[15] The application of log-linear models to 3 or more sources also brings issues regarding model selection and residual model uncertainty as to choice of criteria, where increasing the number of sources exponentially increases the number of possible models that could be fit to the data.[18] With k sources and k-1 interactions, the saturated model will appear optimal by the information criteria; however, caution is advised when choosing a saturated model.[18] Simpler models producing a similar estimate are preferable since a complex unknown k-way interaction in the population being analyzed could produce the same result, which is hard to confirm without independent information.[18] Sparse cell data will also cause wider confidence intervals and a general

instability of the model, which can be corrected for by using correction procedures proposed by various authors. The relative ease of calculating CR methods tends to obscure the relative difficulties mentioned above of model selection and almost always leads to violations in underlying assumptions of the methodology and a biasing of estimates.[12]

In summary, CR methods involve estimating the number of cases in a defined population using multiple sources of information and provide researchers with an efficient, cost-effective alternative to surveillance which can be inefficient, expensive, and often impossible to conduct.[35] While CHD prevalence at birth has been robustly estimated,[4] population-data on CHD prevalence beyond childhood is a significant knowledge gap in the public health community.[1] CR methods have been employed to estimate the prevalence of a wide range of medical conditions including diabetes, cancers, HIV, stroke, inflammatory bowel disease, Tuberculosis, to name a few and is becoming increasingly popular to estimate "hidden populations" such as the homeless or prostitutes.[35] To avoid intensive data collection processes, CR methods allow for leveraging of existing databases, linkage of these data sources, and subsequent de-duplication of these various sources to allow for a reasonable population-based estimate of CHD cases across the lifespan.[1] For this method to be as accurate as possible, it is necessary to clearly define the source population and to maximize case ascertainment by using multiple data sources such as clinical records, hospital discharge records, and insurance databases. Using this approach, previous barriers to estimation including mortality of the birth cohort and late CHD diagnosis not apparent at infancy can be accounted.[1] Common limitations to CR method with application to human populations

include the assumption of equal ascertainment probability between sources due to unidentified covariates (location in study area, severity of disease, human behavior, gender, etc.), as well as the assumption of a closed population which is likely violated due to fatality of disease.[36] Due to the fragmented US healthcare system and lack of CHD surveillance beyond infancy, CR methods allow for the better characterization of the CHD burden across the lifespan.

CHAPTER II: Manuscript

USING CAPTURE-RECAPTURE METHODOLOGY TO ESTIMATE ADOLESCENT AND ADULT CONGENITAL HEART DEFECT (CHD) PREVALENCE IN FIVE METROPOLITAN GEORGIA COUNTIES: 2008-2010

Katherine John

Introduction

Congenital Heart Defects (CHD) are the most common types of birth defect, yet little is known about the public health impact of CHD prevalence across the lifespan.[1] CHDs include birth defects of the heart and/or primary vessels connected to the heart, and in administrative records are coded separately for each abnormality of structure or blood flow, so that an individual can have multiple codes to describe one CHD.[3] Advances over the past four decades in the diagnosis and treatment of children with CHD have resulted in greater than 85% survival of this population into adulthood.[5] Most CHD patients require lifelong cardiology care with roughly half of this population recommended by published guidelines to seek care from cardiac specialists.[5]

While estimates of CHD prevalence at birth have been conducted using birth certificates and hospital birth records in multiple studies, estimates of CHD prevalence beyond childhood in the United States have not been conducted.[1] Estimates of total and age-specific CHD prevalence across the lifespan would allow better characterization of the disease burden of morbidity, mortality, healthcare use, healthcare cost, disability, and non-CHD attributable costs.[1] In 2012, the Centers for Disease Control and Prevention

(CDC) received funding from Congress to enhance and expand public health tracking to improve understanding of CHD across the lifespan. With this funding, the CDC is collaboratively working with Emory University, the Massachusetts Department of Public Health, and New York State Department of Health to pilot a population-based surveillance system of CHDs among adolescents and adults.[37] All three sites are using similar criteria to identify individuals with CHDs, and a 16-member External Guidance Committee consisting of medical and birth defects monitoring experts has been established to provide input on the planning and progression of the project.[37] This analysis is being conducted as part of Emory University's Cooperative Agreement with the CDC on this pilot surveillance effort. The research aim of the current study is to determine the Congenital Heart Defect (CHD) prevalence in five metropolitan counties in the state of Georgia among residents aged 11-64 years from 2008-2010 using capture-recapture (CR) methodology. The data sources were chosen to capture approximately 90% of CHD cases who had sought care within a 3-year time span (2008-2010) and who were at least 11 years of age and not older than 64 years of age and resident in either Clayton, Cobb, DeKalb, Fulton, and Gwinnett counties in Georgia as of 1/1/2010.

Hypotheses

Among the adolescent data sources, positive dependencies will be revealed, while lack of dependence or minimal negative dependence will be seen in the adult data sources.

CR methods using Poisson modeling will yield larger estimates of missing adult CHD cases than seen in the adolescent CHD cases during the period from January 1, 2008 to December 31, 2010.

Methods

Data Sources

CHD cases were collected from seven data sources believed to capture approximately 90% of those patients who sought healthcare in the state of Georgia between 2008 and 2010: Emory Healthcare, St. Joseph's Hospital, Grady Health, the Sibley Heart Center, Pediatric Cardiology Services (PCS), Children's Health Care of Atlanta (CHOA), and Georgia Medicaid claims. Possible reasons for the approximately ten percent of CHD cases not captured include that these individuals sought medical care elsewhere.[39] An unknown percentage of adolescents and adults living with CHD might not have obtained a CHD-related healthcare visit from 2008 through 2010. These "missed" cases might include CHD patients with less severe CHD conditions or defects that have spontaneously closed. Both billing records and medical/clinical records were obtained from the following clinical sites: Emory Healthcare, St. Joseph's Hospital, Grady Health, CHOA, Sibley Heart Center, and PCS. Georgia Medicaid administrative claims data for individuals with a CHD diagnosis were obtained from the Centers for Medicare and Medicaid Services (CMS) via Research Data Assistance Center (ResDAC), a CMS contractor which assists academic, government, non-profits and for-profits. Adolescents were defined as patients 11-20 years of age, and adults were defined as patients 21-64 years of age. The Sibley, CHOA, and PCS data sources provided adolescent CHD data, while the other sources contain the adult CHD population with the exception of Medicaid which contained both adolescents and adults cohort.

Emory Healthcare and St. Joseph's Hospital

Emory Healthcare is the largest health care system in the state of Georgia encompassing a multitude of hospitals, clinics, and local practices.[40] St. Joseph's Hospital, founded in 1880, is Atlanta's longest-serving hospital. In 2012, Emory Healthcare and St. Joseph's Hospital partnered. During the time frame of this study, St. Joseph's Hospital remained a separate independent healthcare facility from Emory Healthcare, and the CHD data obtained from these two entities came to us separately as such. Today, the Emory-St. Joseph Heart and Vascular Institute serves as one of the few heart transplant centers in the state of Georgia and is one of the largest and most decorated cardiac care programs in the country.[40]

Grady Health System

Grady Health System is one of the region's premier level 1 Trauma Centers committed to improving the health and quality of comprehensive healthcare to underserved individuals living in Fulton and DeKalb counties and other metro-Atlanta counties and the entire state of Georgia. Grady Health manages approximately 600,000 patients/year with a majority being enrolled in either Medicare or Medicaid.[41] Grady Health consists of eight facilities located in the surrounding Atlanta area with its Cardiac Clinic housed at the main Grady Memorial Hospital located in downtown Atlanta. This clinic provides comprehensive cardiac care for a variety of conditions and diseases.[41]

Children's Healthcare of Atlanta (CHOA) and Sibley Heart Center

Children's Healthcare of Atlanta (CHOA) consists of pediatric facilities across the state of Georgia dedicated to treating and providing care to children and adolescents.

Ranked 4th in the country by U.S. News & World Report, Sibley Heart Center is one of the top pediatric cardiac programs with 20 outpatient locations and 40 hospitals in the state of Georgia.[42] Sibley offers a spectrum of cardiac programs and services from birth until the age of 21 and has multidisciplinary teams offering specialized care designed especially for children and adolescents who need treatment and management of cardiovascular conditions.

Pediatric Cardiology Services

Pediatric Cardiology Services (PCS) is a group of certified Pediatric Cardiologists who specialize in the care of infants, children, and adolescents in need of high-risk cardiac care.[43] There are currently six locations across the state of Georgia that provide comprehensive cardiac services including cardiac evaluation with diagnostic equipment, prevention counseling, and management of cardiac problems. PCS cardiologists work closely with local neonatologists and primary care physicians to provide the high quality diagnostic and supportive services for patients and families.[43]

Medicaid

Medicaid is a social health care program for families and individuals with low income and resources. The state and federal governments jointly fund the program, with each state having its own criteria for determining eligibility into the program based on state demographics and geography. In Georgia, the Medicaid program provides health care for more than 600,000 residents with low incomes including children, pregnant women, the disabled, and the blind.[39] Disability claims for persons with CHDs can be made to the Social Security Administration (SSA) if they meet the general disability

requirements and qualify for symptomatic congenital heart disease.[44] Even if they do not qualify for one of the listing requirements for symptomatic congenital heart disease, they may still be approved for disability through a physical residual functioning capacity assessment, through the opinion of a licensed physician, and through evidence of emotional/psychological impairment or complications.[44] The primary source for the Medicaid cohort comes from the Medicaid Statistical Information System (MSIS) from which Medicaid Analytic eXtract (MAX) files are constructed.[45] Medicaid data include eligibility status, demographics, claims histories with diagnosis codes, procedure codes, and dates of service. Medicaid data are obtained strictly from billing records, and so, this source is considered to be solely administrative in nature.

Additional Sensitivity Testing Using Data from the MACDP

The Metropolitan Atlanta Congenital Defect Project (MACDP) is a population-based surveillance system for birth defects that was established in 1967 by the Centers for Disease Control and Prevention (CDC), Emory University, and the Georgia Mental Health Insititue.[3] Up until recently, the MACDP conducted surveillance through active case-finding and multiple-source case ascertainment in the five counties in metropolitan Atlanta included in this analysis.[4] The purpose of the MACDP was to provide early warning of increases in the prevalence of birth defects by monitoring trends over time; however, this has evolved to include the monitoring of births for any unusual patterns suggestive of environmental influences, development of a case registry for use in epidemiological and genetic studies, quantifications of morbidity and mortality with birth defects, provision of data for health policy and educational purposes, and to provide public health training in surveillance and epidemiological methods.[3] Cases include those

births where the mother was a resident of one of the counties, where the fetus had a major structural or chromosomal defect present at birth that adversely affects health or development, where the infant, fetus or child was at least 20 weeks gestation at the time of delivery, and where the defect was diagnosed before the child's 6th birthday.[3] From the MACDP cohort, CHD cases were extracted according to the ICD9s included in the case definition of the larger CDC pilot project (see Appendix A). For this analysis, data from the MACDP form a comparison with prevalence estimates using CR methodology.

Data Collection

Demographic and encounter level data were obtained for males and females with a CHD diagnosis, who were at least 11 years of age by January 1, 2010 and not older than 64 years of age, living in the state of Georgia (see Appendix A for case definition), and who sought healthcare either for their CHD condition or otherwise between January 1, 2008 to December 31, 2010. All data obtained were cleaned and de-duplicated within data source. For those datasets which had last name, first name, date of birth, and gender, those fields were combined and a unique case ID was created to determine duplicate cases within data source. In addition, most datasets had an internal unique identifier which was also used to de-duplicate records. Data sources were then linked by the unique case ID and tracked across data sources in a 'Master' Microsoft Access database table.

Linking Across Sources

	Last Name	First Name	DOB	Gender	SSN
EMORY	X	X	X	X	X
ST. JOE'S	X	X	X	X	X
GRADY	X	X	X	X	X
MEDICAID	not available	not available	X	X	X
CHOA	X	X	X	X	X
SIBLEY	X	X	X	X	X
PCS	X	X	X	X	not available
MACDP	X	unreliable	X	X	not available

Statistical Methods

The data were evaluated through a two-source CR method. Using the number of unique individuals captured in each of two data sources and the number of unique individuals captured in both those data sources, the number of individuals in the total population and the number of individuals missed were estimated. Given that: 1) the population was closed during the capture time period (meaning that no individuals were lost); 2) each person was matched from capture (source 1) to recapture (source 2); and 3) for each data source, the individual had the same probability to be included (indicating that the two sources are independent form one another), accurate estimates using this two-source CR method were calculated.[13]

The structure of the two-source CR data analysis is below (Table 1).

Table 1. Two-Source Contingency Table for Capture Recapture (CR) Method

		Source 1		
		Yes	**No**	
Source 2	**Yes**	X_{12}	X_2	N_2
	No	X_1	X_0	
		N_1		N

Where:

- N = all cases occurring (estimated);
- N_1 = Total cases in Source 1;
- N_2 = Total cases in Source 2;
- X_{12} = cases found by both Source 1 and Source 2;
- X_1 = cases found in Source 1, but not Source 2;
- X_2 = cases found in Source 2, but not Source 1; and
- X_0 = cases not found in Source 1 or Source 2.

To observe the number of number of cases between the two sources, the number of cases found in both sources (overlapping) is summed with the number of individuals found exclusively in Source 1 and the number of individuals found exclusively in Source 2 (Eq. 1).[46] To assess the contributions from each source, two-source CR was conducted using each source combination. For eight sources, this resulted in 28 different two-source estimates. In two source capture recapture situations to estimate population size, the Lincoln-Peterson estimator, based on the odds ratio, is most commonly used with the assumption that identifying sources are independent and that cases are equally likely to

be identified in each source. Another form of this estimator, the Chapman estimator, was used for this study since it has been found to have optimal properties under a wide range of conditions and is less affected by small sample bias (zeros found in the table) resulting in a nearly "unbiased estimate".[12,36] Contributions of cases from the two sources used in the following formulae can be found using the table above (Table 1). The estimated total number of cases of CHD, using Chapman's modified estimate of the Lincoln-Peterson method in two-source CR, was calculated using the total number of cases identified in Source 1 plus one (to eliminate small sample bias), the total number of cases identified in Source 2 plus one, and the number of overlapping CHD cases between the two sources (Eq. 2).[46] The variance for the total population size estimate using the Chapman estimator was calculated using the below formula (Eq. 3).[46] Further, the number of CHD cases not identified by either source was estimated using either of the below formulas (Eq. 4 or 4a).[46] These formulas provide similar estimates; however, equation 4a was reported to coincide with the estimates of total population size reporting. This method is particularly useful for two-source capture analysis when pooling lists.

$$\text{Eq. 1.} \quad \text{Observed N} = X_{12} + X_1 + X_2$$

$$\text{Eq. 2.} \quad \text{Estimated } \widehat{N} = \frac{(N_1 + 1)(N_2 + 1)}{X_{12} + 1}$$

$$\text{Eq. 3.} \quad \text{Variance N} = \frac{(N_1 + 1)(N_2 + 1)(X_1)(X_2)}{({X_{12}}^2)(X_{12} + 2)}$$

$$\text{Eq. 4.} \quad \text{Estimated Missing } X_0 = \frac{(X_1)(X_2)}{X_{12}}$$

$$\text{Eq. 4a.} \quad \text{Estimated Missing } X_0 = \widehat{N} - X_1 - X_2 - X_{12}$$

Independence among sources is a major assumption when using two-source CR. Lack of independence could lead to bias in the estimates. In order to check the dependence between sources, the probability of being captured in the Source 1 is compared to the probability of being recaptured in the Source 2. The probability of being captured in Source 1 is estimated as the number of captures in the Source 1 divided by the number of estimated total captures using the CR equation (Eq. 5). The probability of being recaptured is estimated using the number of captures found in both sources divided by the number of captures found in the Source 2 (Eq. 6). When the two sources are independent, the recapture rate is approximately equal to the capture rate in the population ($N_1/N = X_{12}/N_2$).[17] When the two populations are positively dependent, the recapture rate is expected to be larger than the capture rate, thus underestimating the estimated total (X_{12}/N_2 (recapture) > N_1/N (capture)), and when the two populations are negatively dependent, the recapture rate is expected to be smaller than the capture rate leading to an overestimate using the CR method (X_{12}/N_2 (recapture) < N_1/N (capture)).[17]

$$\text{Eq. 5. Capture Rate} = \frac{N_1}{\hat{N}}$$

$$\text{Eq. 6. Recapture Rate} = \frac{X_{12}}{N_2}$$

The term *Source* denotes a list of cases sometimes with or without a unifying characteristic to how they were ascertained into the same list.[12] Any application of CR methodology requires a critical decision on how to define analytic Sources. Different lists may be pooled to derive a larger group at the flexibility and judgment of the investigator.[12]

With the inclusion of MACDP (the Metropolitan Atlanta Congenital Defects Program) and previous prevalence estimates which only utilized cases residing from Clayton, Cobb, Fulton, DeKalb, and Gwinnett counties,[4] it was decided to limit the other seven sources to cases residing in these 5 counties. Upon reviewing how the data sources were constructed, a decision was made to combine data sources based on age. CHOA, Sibley, and PCS data were acquired because they are considered adolescent (pediatric) healthcare providers, while Emory, St. Joseph's, and Grady were targeted because these facilities primarily provide care to adult CHD patients. As such, the individual adolescent data sources were pooled to form an adolescent database and the individual adult data sources were pooled to create an adult database. Before pooling occurred, age outliers were identified and removed. In the "adolescent" data sources, patients 21 years old and older were identified and removed so that when these adolescent sources were pooled, an adolescent cohort between 11 and 20 years old remained. In the "adult" data sources, patients younger than 21 years old were identified and removed so that when these sources were pooled, an adult cohort between 21 and 64 years of age remained.

The pooled dataset had no missing observations for age and consisted of 4,797 individuals, including multiple observations for some individuals. Datasets included a unique de-duplication ID, patient's last name (except for Medicaid), first name (except for Medicaid), date of birth, gender, social security number when available, a unique ID identifier which also served to flag presence in each of the data sources, and a count variable indicating the number of inclusions/captures across datasets for each individual.

Using these pooled sources, separate age-based CR prevalence estimates were conducted. The first analysis focused only on the adolescent population, defined as: 1)

those individuals captured in one of the adolescent datasets; and 2) those adolescents identified in the GA Medicaid claims data who were 11-20 years old. These data were classified into three sources: 1) *Combined CHOA and PCS* (CHD adolescents aged 11-20 found in CHOA and PCS); 2) *Sibley* (adolescents aged 11-20 found in Sibley); and 3) *Adolescent administrative* (CHD Medicaid adolescents age 11-20). CHOA and PCS were combined due to the low number of CHD cases found in PCS; however, to ensure accuracy, Sibley and PCS were also analyzed as a combined source. There were no significantly different results in the estimates and due to the large number of CHD cases in Sibley it was decided to combine CHOA and PCS. The second analysis focused only on the adult population, defined as those CHD patients age 21-64 years. For this analysis the data were classified into four sources: 1) *Adult Emory* (CHD patients aged 21-64 captured in Emory); 2) *Adult Grady* (CHD patients aged 21-64 captured in Grady); 3) *Adult St. Joe's* (CHD patients aged 21-64 captured in St. Joseph's); and 4) *Adult administrative* (CHD Medicaid patients age 21-64 years). The same formulas were used for both of these 2 two-source CR analyses; however, only 1 estimate for total CHD population and those missing was calculated for each age population.

In manual calculations, each of the three adolescent sources was used for 2-source CR calculations; likewise, 2-source CR calculations were conducted using the four adult sources. To check estimates from the manual calculations, logistic modeling was conducted. Modeling also allowed for an easier inclusion of multiple sources. However, when there are data from more than two sources, analysis becomes more complex. As the number of k sources increases, so does the number of estimates that can be derived using the combination of the sources excluding k-1 source interaction.[12] Since

dependencies often exist between two or more sources, log-linear modeling using statistical software was employed incorporating dependencies with interaction terms.[47]

Poisson modeling using the three adolescent and four adult cohort databases was conducted to obtain separate adolescent and adult CHD prevalence estimates and to ensure accuracy in combining the adolescent and adult sources. Poisson regression using the PROC GENMOD procedure was chosen for modeling counts. The Poisson approach is a good fit for CR analyses because it is appropriate when: (1) the captures can be counted in whole numbers; (2) the capture sources are independent from one another; and (3) it is possible to count how many captures have occurred.[48] In order to derive an appropriate estimate, values from the likelihood ratio statistic (Eq. 7), Akaike Information Criteria (AIC) (Eq. 8), and deviances/degrees of freedom (df) determined model fit. The lower the likelihood ratio statistic and AIC, and the closer the value of the deviance/df is to 1, the better the model fit.[12] Once the best estimate of the number of CHD cases was determined, the prevalence of CHD among adolescents in the 5-county area was calculated by dividing the population aged 10-19 as of the U.S. Census of 2010 by the estimated number of adolescent CHD cases aged 11-20 and multiplying by 1,000.[49] Similarly, the prevalence of CHD among adults aged 21-64 was calculated by dividing the population aged 20-64 as of the U.S. Census of 2010 by the estimated number of adult CHD cases, multiplied by 1,000.[49]

All data were analyzed using SAS 9.4 (SAS Institute Inc. North Carolina).

$$\text{Eq. 7. } G^2 = -2\sum \text{Obs}_j \log\left(\frac{Obs_j}{Exp_{ji}}\right) \text{ where j denotes the cell and i denotes the model}$$

$$\text{Eq. 8. } \text{AIC} = G^2 - 2(df)$$

Results

The majority of cases were below the age of 41 (64.1%), female (53.4%), and did not have a race identified (86%) (Table 1). The number of individuals captured by provider source varied in size (range 11-2,556 CHD cases), percent male (range 36%-62%) and mean age (range 18-48 years) (Table 2). Five hundred adult cases were found in adolescent exclusive data sites, while 299 adolescent cases were found in adult exclusive data sites (Table 2). These groups were excluded from further CR calculations.

Of the 4,797 unique CHD cases identified in the five counties, the majority of cases were captured using a single source (76.7%) (Table 3). As the number of capture sources increased, the percentage of captured individuals decreased, with only 17 unique CHD cases captured in 5 of the 8 sources, including the MACDP (Table 3). Age and gender were available for all unique CHD cases captured, while race was identified in only 14% of the total CHD cases (data not shown).

Adolescent clinical sites (n=1,788) (CHOA, Sibley, PCS) and the adolescent Georgia Medicaid population (n=70) were used in manual calculations of two source CR methodology. The estimated CHD population ages 11-20 years ranged from 600-1,903 individuals (Table 4). Tests for dependence between the three adolescent data sources resulted in no dependencies between each of the three CR calculations using two sources (Table 5). The two source CR method was also manually conducted for the adult CHD population using adult clinical and billing sources (n=2,878) (Emory, Grady, St. Joseph's) and adult Georgia Medicaid administrative claims data (n=305). For the six models, the estimated number of CHD cases aged 21-64 years ranged from 1,560-86,769 (Table 4). Each of the four sources was found to be independent of one another (Table

5). Model 3 in the manual CR adolescent analysis and Models 3 and 6 in the manual adult CR methods produced estimated total CHD populations less than the observed cases in each dataset (Table 4).

Multiple Poisson models using combined CHOA/PCS, Sibley, and adolescent Medicaid were conducted using interaction terms to control for possible dependencies between the three sources. Model 4 in the adolescent Poisson calculations produced the lowest AIC criterion and one of the highest likelihood ratio statistics. This model controls for dependency between combined CHOA/PCS and Medicaid and estimates the number of missing CHD cases to be 1,860, while the total CHD adolescent population aged 11-20 years was estimated to be 3,718 (95%CI 3,471-4,004) (Table 6). Multiple Poisson models controlling for dependencies with interaction terms were run for the four adult data sources: Emory, Grady, St. Joseph's, and adult Medicaid. Model 4 in the adult Poisson modeling, which controlled for dependency between Emory and Medicaid, was chosen as the final model. This model estimates the number of missing adult CHD cases to be 12,969 and the total adult CHD population to be 16,152 (95%CI 13,873-18,915) (Table 7).

The final two models for the separate adolescent and adult CR Poisson Modeling analysis were as follows:

$$\text{Adolescent CR Model: Logit P(count)} = \propto + (\text{Sibley}) + (\text{CHOA/PCS}) + (\text{Medicaid}) + (\text{CHOA\&PCS} * \text{Medicaid}), \textit{where distribution} = \textit{Poisson}$$

$$\text{Adult CR Model: Logit P(count)} = \propto + (\text{Emory}) + (\text{St Joe}) + (\text{Grady}) + (\text{Medicaid}) + (\text{Emory} * \text{Medicaid}), \textit{where distribution} = \textit{Poisson}$$

Using the estimates from these models, the adolescent prevalence was estimated to be 7.85 per 1,000 residents aged 10-19 in 2010, and 6.08 per 1,000 residents aged 20-64 in the 5-county metro Atlanta area. In a supplemental analysis which retained those cases whose age fell outside the mission of the data source, meaning adults found in the adolescent data sources or adolescents found in the adult data sources, CHD population estimates increased by roughly 1,500 and 1,000 in the adolescent and adult populations, respectively. Of the 5,271 MACDP cases, 2,186 were adolescents as of January 1, 2010, and 3,085 cases were adults, aged 21-42 (since births prior to 1967 occurred prior to the initiation of the MACDP). This is a smaller number than the estimated CHD cases in comparable age groups now residing in the 5-county metropolitan Atlanta area (3,718 aged 11-20 and 6,210 aged 21-42, respectively).

Discussion

The aim of this study was to use electronic medical and billing records, and administrative claims data to estimate the prevalence of the adolescent and adult CHD population residing in the 5-county metropolitan Atlanta area from January 1, 2008 to December 31, 2010. To assess whether these estimates were reasonable, it was necessary to evaluate the possible source dependencies that existed among the various databases acquired. While results from the adolescent two source CR analyses revealed capture and recapture rates within 0.5% of one another (reflecting independence among data sources) did not support the hypothesis, the adult two source CR analyses which also reflected independence, did support the hypothesis.

Using warehouse data from the Georgia Department of Public Health, the population in 2010 of individuals residing in the five metropolitan Atlanta counties was 473,533 among persons aged 10-19 years and 2,133,575 among persons 20-64 years.[49] Congenital birth defects have been estimated to affect 1% of births,[1] leading to an adolescent CHD population consisting of 4,735 persons and an adult population of 21,336 persons during this time period and residing in these five counties. Use of adolescent CR methods provided a close approximation with 3,718, with an estimated prevalence of 0.785 percent. Adult CR analyses yielded a prevalence estimate of 0.608 percent, which also rounds to 1% and may fairly represent the adult CHD population who were born at a time when survival rates were lower than contemporary survival rates. Consistent with the hypothesis regarding healthcare utilization, the missing adult CHD case population was estimated to be roughly 4 times the size as the estimated missing adolescent CHD cases.

For all CR calculations, the Chapman estimator method was chosen due to optimal properties under a wide range of conditions.[12] The Chapman estimator is often employed for use with small cells; however, it can be appropriately used for large data as seen in this study. A major assumption in use of the CR method is the consistency of pre-specified analytic sources to be used. Age of the patient was anticipated to affect the probability of capture in the sources and as such, the decision to stratify into two age categories was conducted. It was clear from restricting age to site specifications, based on the mission statement and services available by data source (site), that adolescent clinical sites contained a large portion of outlying adult CHD patients. In other words, there are a large number of patients seen by pediatric cardiology care providers who are

old enough to have already transitioned into adult care for their CHD, but who remain under the care of their pediatric provider. This retention of adult CHD patients could be due to institutional problems contributing to a lack of outpatient age restriction, barriers to adult care, or referral problems.[50,51] In limiting the CR method to the mission of the source, the number of captures in each site was confined which could have affected estimates of prevalence of each population. Outlying adolescent CHD cases were also found in the adult sites; however, this was not as frequent as adults found in the adolescent specific sites.

Both the adolescent and the adult sources were found to be independent of the other sources contained in both CR analyses. This independence could be due to possible confounding variables not investigated thereby yielding an overestimate of prevalence. Biased estimates of population size could arise due to variability in the sample populations utilized. It is an assumption of this method that the same population is being sampled between the first capture and the recapture. Population characteristics influenced by geographic location, immigration, and missing demographics, can make data sources appear independent from one another, when, in fact, they may be divergent from one another.[34] Even when assumptions are violated, determination of the direction of dependency can predict if the capture-recapture estimate is likely plausible or an over- or under-estimate of the true total.[12]

The larger numbers of missing adult CHD patients could be due to early mortality; however, severity of disease was not included in this analysis. Despite advances in diagnostic procedures, early mortality is an outcome for people with CHDs, especially for those classified as complex or severe. For example, one clinic study found

the mean age at death for patients with moderate to complex CHD to be 37 years.[52] For patients with mild forms of CHD such as aortic and mitral valve disease, the rise in prevalence throughout the lifespan is consistent with improvement in diagnostic techniques and/or presentation in adulthood, with an increased likelihood of being captured with longer observation periods.[8]

The first year of life is the most critical period of survival for an infant with a CHD, after which survival probabilities for infants and children vary by type of CHD with up to 8 year survival ranging from 50-85%.[53,54] It is therefore somewhat surprising that the estimates of adolescents and adults living with CHD, based on diagnosis at birth up to 6 years of age from the MACDP, are considerably less than the numbers observed and the total estimated numbers from CR analyses. The population of Atlanta has grown dramatically over the last few decades (more than doubling in the 5 counties between 1970 and 2000).[55] Thus, in-migration of CHD survivors is one likely explanation for the discrepancy. On the other hand, only 12% of the MACDP cases were located in clinical and billing, or administrative records from 2008 – 2010, suggesting an under-ascertainment of CHD cases owing to lack of healthcare seeking or moving out of the area. A search of a sample of CHD cases captured by the MACDP who were not recaptured by any other data source was conducted to determine if these cases still resided in the state of Georgia. Preliminary results revealed that while 19% of cases had moved out of state, 48% remained in state with 68% of those still residing in the 5 county MACDP catchment area.

One limitation of this study was in the construction of the unique identifier used and the resulting matching. Deterministic (exact) matching using last name, first name,

social security number, date of birth, and gender could be utilized for seven of the eight sources, including MACDP used in the supplemental analysis. Medicaid, derived from billing data, lacked information that was found in the other sites. For this reason, probabilistic matching was used for Medicaid which could have limited the number of unique captures from this source. Another limitation to this study is lack of consistent reporting of age and marital status. While marital status does not apply mostly to the younger age groups, both of these variables are associated with CHD ascertainment and could have been controlled for in modeling. Due to the incompleteness of marital status, this variable was omitted.

CR methods assume that each case was truly captured correctly; in other words, that the case is a true case. A sample Medical record review of ICD-9 code 745.5 for CHOA, Sibley and Emory patients showed that 47% were confirmed as ASD, while 53% were misclassified; of those misclassified, 15% confirmed "normal." In another validation study looking at Emory Healthcare patients with a VSD in isolation who were greater than 40 years old found that 78% had a confirmed VSD, while 22% were misclassified; of those misclassified; over 50% had either coronary artery disease or a post Myocardial Infarction condition. Lastly, two other validation studies were conducted among Emory Healthcare patients who were greater than 40 years old, one looking at ICD-9 code 746.85 in isolation, which is coronary artery anomaly, and the other looking at 746.9 in isolation, which is unspecified congenital anomaly of the heart. These assessments found that 95% had a confirmed coronary artery anomaly with only 5% misclassified, while 76% were confirmed to have an unspecified congenital heart anomaly with 24% misclassified. Apparently, there is a fair amount of misclassification

in case ascertainment which could lead to an overestimation of the prevalence of CHD without careful scrutiny of the CHD diagnoses being assigned.

REFERENCES

1. Oster ME, Riehle-Colarusso T, Simeone RM, et al. Public Health Science Agenda for Congenital Heart Defects: Report from a Centers for Disease Control and Prevention Experts Meeting. *Journal of the American Heart Association.* 2013:1-10.
2. US Department of Health and Human Services. What are Congenital Heart Defects? *Health Information for the Public* 2014. <www.nhlbi.nih.gov/health/health-topics/topics/chd/>
3. Correa A, Cragan JD, Kucik JE, et al. Metropolitan Atlanta Congenital Defects Program: Reporting Birth Defects Surveillance Data 1968-2003. *Birth Defects Research* 2007;79:65-93.
4. Reller MD, Strickland MJ, Riehle-Colarusso T, Mahle WT, Correa MD. Prevalence of Congenital Heart Defects in Metropolitan Atlanta, 1998-2005. *The Journal of Pediatrics.* 2008:807-813.
5. Gurvitz M, Valente AM, Broberg C, et al. Prevalence and Predictors of Gaps in Care Among Adult Congenital Heart Disease Patients. *Journal of the American College of Cardiology.* 2013;61(21):1-5.
6. Pillutla P, Shetty KD, Foster E. Mortality Associated With Adult Congenital Heart Disease: Trends in the US Population from 1979 to 2005. *American Heart Journal.* 2009;158(5):874-879.
7. Russo C, Elixhauser A. Hospitalizations for Birth Defects, 2004. HCUP Statistical Brief #24. 2007;

<http://www.hcupus.ahrq.gov/reports/statbriefs/sb24.pdf.> Accessed January 2015.

8. Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime Prevalence of Congenital Heart Disease in the General Population from 2000 to 2010. *Circulation.* 2014:1-24.
9. Congenital Heart Public Health Consortium. Fact Sheet. 2011. <www.chphc.org>
10. O'Leary JM, Siddiqi OK, Ferranti S, Landzberg M, Opotowsky A. The Changing Demographics of Congenital Heart Disease Hospitalizations in the United States 1998 Through 2010. *JAMA.* 2013;309(10):984-986.
11. US Department of Health & Human Services. How the Health Care Law is Making a Difference for the People of Georgia. *Health Care* 2014. <http://www.hhs.gov/healthcare/facts/bystate/ga.html>
12. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev.* 1995;17(2):243-264.
13. Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. *Am J Epidemiol.* 1995;142(10):1047-1058.
14. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis.* 1974;27(1):25-36.
15. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol.* 1999;52(10):917-926; discussion 929-933.

16. Evans MA, Bonett DG, McDonald LL. A general theory for modeling capture-recapture data from a closed population. *Biometrics.* 1994;50(2):396-405.

17. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med.* 2001;20(20):3123-3157.

18. Hook EB, Regal RR. Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *Am J Epidemiol.* 1997;145(12):1138-1144.

19. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Stat Med.* 1984;3(3):287-291.

20. Sekar CC, Deming WE. On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association.* 1949;44:101-115.

21. Herzog TN. Applications of Capture-Recapture Methods. 40th Actuarial Research Conference2006:1-11.

22. Verstraeten T, Baughman AL, Cadwell B, et al. Enhancing Vaccine Safety Surveillance: A Capture-Recapture Analysis of Intussecption after Rotovirus Vaccination. *American Journal of Epidemiology.* 2001;154(11):1006-1012.

23. McGilchrist CA, McDonnell LF, Jorm LR, Patel MS. Loglinear Models Using Capture-Recapture Methods to Estimate the Size of a Measles Epidemic. *Journal of Clinical Epidemiology.* 1996;49(3):293-296.

24. LaPorte RE, Dearwater SR, Chang YF, et al. Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. *Am J Epidemiol.* 1995;142(10):1069-1077.

25. Laska E. Editorial: The Use of Capture-Recapture Methods in Public Health. *Bulletin of the World Health Organization.* 2002;80 (11):845.

26. Tilling K. Capture-recapture methods--useful or misleading? *Int J Epidemiol.* 2001;30(1):12-14.

27. US Department of Health and Human Services. NIH Clinical Research Trials and You: List of Registries. 2014. <http://www.nih.gov/health/clinicaltrials/registries.htm>

28. Civic Impulse L. S. 1382 (110th): ALS Registry Act. 2008. Accessed August 30, 2014, 2014. < https://www.govtrack.us/congress/bills/110/s1382>

29. Wittie M, Nelson LM, Usher S, Ward K, Benatar M. Utility of capture-recapture methodology to assess completeness of amyotrophic lateral sclerosis case ascertainment. *Neuroepidemiology.* 2013;40(2):133-141.

30. Corrao G, Bagnardi V, Vittadini G, Favilli S. Capture-recapture methods to size alcohol related problems in a population. *J Epidemiol Community Health.* 2000;54(8):603-610.

31. Bruno G, LaPorte RE, Merletti F, Biggeri A, McCarty D, Pagano G. National diabetes programs. Application of capture-recapture to count diabetes? *Diabetes Care.* 1994;17(6):548-556.

32. La Ruche G, Dejour-Salamanca D, Bernillon P, et al. Capture-recapture method for estimating annual incidence of imported dengue, France, 2007-2010. *Emerg Infect Dis.* 2013;19(11):1740-1748.

33. Somers EC, Marder W, Cagnoli P, et al. Population-Based Incidence and Prevalence of Systemic Lupus Erythematosus. *Arthritis & Rheumatology*. 2014;66(2):369-378.

34. Stephen C. Capture-recapture methods in epidemiological studies. *Infect Control Hosp Epidemiol.* 1996;17(4):262-266.

35. Morrison A, Stone DH. Capture-recapture: a useful methodological tool for counting traffic related injuries? *Injury Prevention.* 2000;6:299-304.

36. Brittain S, Bohning D. Estimators in capture-recapture studies with two sources. *Advanced Statistical Analyses.* 2009;93:23-47.

37. Divison of Birth Defects and Developmental Disabilities. Population-Based Surveillance of Congenital Heart Defects among Adolescents and Adults. 2014. <http://www.cdc.gov/ncbddd/heartdefects/documents/chdsurveillance_factsheet_cleared.pdf>

38. Honein MA, Paulozzi LJ. Birth Defects Surveillence: assessing the "gold standard". *American Journal of Public Health.* 1999;89(8):1238-1240.

39. Book W, Raskind-Hood C, Hogue C. CDC FOA dd12-1207. Atlanta, GA: CDC; 2014.

40. Emory Healthcare. About Emory Healthcare. 2014. <http://www.emoryhealthcare.org/about-us/index.html> Accessed August 20, 2014.

41. Grady Healthcare System. Learn About Us. 2014. <http://gradyhealth.org/learn-about-us>

42. Children's Healthcare of Atlanta. Children's Healthcare of Atlanta Sibley Heart Center. 2014. <http://www.choa.org/heart>

43. Pediatric Cardiology Services. Pediatric Cardiology Services. 2012. <http://www.pediatriccardiologyservices.com> Accessed August 30, 2014, 2014.>

44. Linebaugh M. Congenital Heart Disease: When Are Disability Benefits Available? *Disability Secrets* 2015. <http://www.disabilitysecrets.com/resources/social-security-disability-coverage/congenital-heart-disease.htm> Accessed April 6, 2015, 2015.

45. Centers for Medicare and Medicaid. Medicaid Data Sources-General Information. 2014. <http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/index.html>

46. Reintjes R, Termorshuizen F, MJW. VdL. Assessing the Sensitivity of STD Surveillance in the Netherlands: an application of the Capture-Recapture Method. *Epidemiology and Infections.* 1999;122:97-102.

47. Orton H, Rickard R, Gebella B. Epidemiology. *1999.* Capture-Recapture Estimation Using Statistical Software;10(5):563-564.

48. University of Massachusetts Amherst. Statistics: The Poisson Distribution. 2007. <http://www.umass.edu/wsp/resources/poisson/> Accessed February 18, 2014, 2014.

49. OASIS Web Query- Population Statistics. Georgia Department of Public Health and Office of Health Indicators for Planning; 2015. <https://oasis.state.ga.us/oasis/oasis/qryPopulation.aspx> Accessed April 6,2015.

50. Norris MD, Webb G, Drotar D, et al. Prevalence and Patterns of Retention in Cardiac Care in Young Adults with Congenital Heart Disease. *Journal of Pediatrics.* 2013;163(3):902-904.

51. Fernandes SM, Khairy P, Fishman L, et al. Referral Patterns and Perceived Barriers to Adult Congenital Heart Disease Care. *Journal of American College of Cardiology.* 2012;60(23):2411-2418.

52. Reid GJ, Webb GD, Barzel M, McCrindle BW, Irvine JM, Siu SC. Estimates of Life Expectancy by Adolescents and Young Adults with Congenital Heart Disease. *Journal of American College of Cardiology.* 2006;48(2):249-355.

53. Wang Y, Liu G, Canfield MA, et al. Racial/Ethnic Differences in Survival of United States Children with Birth Defects: A Population-Based Study. *Journal of Pediatrics.* 2015:1-10.

54. Moons P, Bovijn L, Budts W, Belmans A, Gewillig M. Temporal Trends in Survival to Adulthood Among Patients Born with Congenital Heart Disease from 1970 to 1992 in Belgium. *Circulation.* 2010;122:2264-2272.

55. US Census Bureau. Population Estimates: Population in the US by County in Georgia (1970-2010). Last revised February 5, 2015. <http://www.google.com/publicdata/explore?ds=kf7tgg1uo9ude_&met_y=population&idim=county:13063:13151&hl=en&dl=en#!ctype=l&strail=false&bcs=d&nselm=h&met_y=population&scale_y=lin&ind_y=false&rdim=country&idim=county:13063:13151:13067:13089:13121:13135&ifdim=country&hl=en_US&dl=en&ind=false> Accessed April 10, 2015.

56 Warnes CA, Williams RG, Bashore TM, et al. ACC/AHA 2008 Guidelines for the Management of Adults with Congenital Heart Disease. *Circulation.* 2008;e714-e820.

57. Heery E, Sheehan AM, While AE, Coyne I. Experiences and Outcomes of Transition from Pediatric to Adult Health Care Services for Young People with Congenital Heart Disease: A Systematic Review. *Congenital Heart Disease.* 2015:1-15.

58. American College of Cardiology Foundation, ACHA. PATCH (Provider Action for Treating Congenital Hearts). 2014. <http://www.achaheart.org/Portals/0/pdf/PATCH/PATCH%20Program%20Addresses.pdf> Accessed February 21, 2015.

59. American Heart Association. Healthcare Reform and You. 2014. <http://www.heart.org/HEARTORG/Advocate/IssuesandCampaigns/Health-Care-Reform-and-You_UCM_314633_Article.jsp> Accessed February21, 2015.

TABLES

Table 1. Sociodemographics of CHD Patients Captured Between January 1, 2008 and December 31, 2010, Limited to Five Metropolitan Atlanta, Georgia Counties (Clayton, Cobb, DeKalb, Fulton, and Gwinnett)

Characteristic	**N (%)**
Overall	4,797 (100%)
Gender	
Male	2,234 (46.6%)
Female	2,563 (53.4%)
Age Group	
11-20	1,621 (33.8%)
21-30	762 (15.9%)
31-40	692 (14.4%)
41-50	646 (13.5%)
51-60	757 (15.8%)
61-64	319 (6.7%)
Race	
American Indian	1 (0.02%)
Asian	26 (0.5%)
Black	231 (4.8%)
Hawaiian	3 (0.06%)
White	425 (8.9%)
Unknown	4,111 (85.7%)

Table 2. Distribution of Age and Gender for Unique Georgia CHD Cases (Ages 11- 64) by Seven Data Sources Limited to Five Counties (Clayton, Cobb, DeKalb, Fulton, and Gwinnett)

SOURCE[1]	Captured N (%)[2]	% Male[3]	Age Mean (SD)	Captured N [11-20 years]	Captured N [21-64 years]
CHOA	511 (10.65%)	46.0%	18.83 (7.01)	394	117
SIBLEY	1,766 (36.81%)	47.5%	18.10 (6.21)	1,384	382
PCS	11 (0.23%)	36.4%	17.09 (5.28)	10	1
EMORY	2,556 (52.28%)	45.9%	40.97 (14.77)	266	2,290
GRADY	341 (7.11%)	40.8%	41.00 (14.29)	31	310
ST. JOSEPH	280 (5.84%)	61.8%	48.34 (11.55)	2	278
MEDICAID	375 (7.82%)	36.0%	35.73 (15.26)	70	305

[1] CHD patients could be captured in multiple data sources.
[2] Percent was calculated as the number of individuals found in a particular source by the total number of observations in the dataset (n=4,797)
[3] Percent male was calculated using total captured in individual data source.

Table 3. Number of Captures for Unique CHD Cases in the Five Metropolitan Atlanta, Georgia Counties*, 11-64 Years (N=4,797)

Number of Captures for a Unique CHD Case	Count of CHD Patients Captured	Percent (%)
1	3,678	76.7%
2	832	17.3%
3	210	4.4%
4	60	1.3%
5	17	0.4%

* Clayton, Cobb, DeKalb, Fulton, Gwinnett

Table 4. Manual CR Analysis of Both the Adolescent and Adult Populations Using Two-Sources in the Five Metropolitan Atlanta, Georgia Counties[1]

Model	Source 1	Source 2	Total Population					
			Source 1 (N1)	Source 2 (N2)	Both (X12)	Est. Missing	CR Est. Total	95% CI Total
Adolescent Two Source								
1	SIBLEY	CHOA/ PCS[2]	1,384	396	288	411	1,903	(1,801, 2,005)
2	SIBLEY	MCAID[3]	1,384	70	52	453	1,855	(1,606, 2,104)
3	CHOA/ PCS[2]	MCAID[3]	396	70	46	180	600	(505, 694)
Adult Two Source								
1	EMORY	STJOE	2,290	278	19	29,411	31,960	(18,157, 45,762)
2	EMORY	GRADY	2,290	310	36	16,693	19,257	(13,398, 25,116)
3	EMORY	MCAID[4]	2,290	305	241	543	2,897	(2,739, 3,055)
4	ST. JOSEPH	GRADY	278	310	0	86,181	86,769	
5	ST. JOSEPH	MCAID[4]	278	305	10	7,188	7,761	(3,113, 12,410)
6	GRADY	MCAID[4]	310	305	60	1,005	1,560	(1,243, 1,877)

[1]Clayton, Cobb, DeKalb, Fulton, Gwinnett
[2]CHOA (n=394) and PCS (n=10) were combined into one source. Eight CHD cases were found in both of these sources leaving 396 total
[3]Medicaid administrative data for adolescents was limited to ages 11-20 years
[4]Medicaid administrative data for adults was limited to ages 21-64 years

Table 5. Dependency Results between Each Two-Source CR Analysis Conducted on the Adolescents (11-20 years) and Adults (21-64 years) in the Five Metropolitan Atlanta, Georgia Counties[1]

Model	Source 1	Source 2	Recapture Rate	Capture Rate	Dependence[2]
Adolescent Models					
1	SIBLEY	CHOA/ PCS	0.727273	0.7274341	Independent
2	SIBLEY	MCAID	0.742857	0.7459399	Independent
3	CHOA/ PCS	MCAID	0.657143	0.6603044	Independent
Adult Models					
1	EMORY	STJOE	0.068345	0.071653	Independent
2	EMORY	GRADY	0.116129	0.118919	Independent
3	EMORY	MCAID	0.790164	0.790505	Independent
4	ST. JOSEPH	GRADY	0.000000	0.003204	Independent
5	ST. JOSEPH	MCAID	0.032787	0.035819	Independent
6	GRADY	MCAID	0.196721	0.198705	Independent

[1]Clayton, Cobb, DeKalb, Fulton, Gwinnett
[2]Dependence was assessed to 0.005

Table 6. CR Analysis Using Poisson Modeling of Three Adolescent Sources to Estimate Missing and Total CHD Cases in Five Metropolitan Atlanta, Georgia Counties[1]

Model	Interaction Terms[2]	AIC Criterion	G^2	Deviance / df	Pred. Missing	95%CI Missing	Est. Total[4]	95%CI Total
1[3]	none	114	12534	19	1,610	1,406-1,843	3,468	3,263-3,701
2	S1*S2	104	12539	23	1,049	784-1,403	2,907	2,642-3,261
3	S1*S3	115	12534	29	1,589	1,367-1,846	3,447	3,225-3,704
4	S2*S3	58	12562	0.3	1,860	1,613-2,146	3,718	3,471-4,004
5	S1*S2*S3	92	12546	17	1,818	1,573-2,102	3,676	3,431-3,960
6	S1*S2 S1*S2*S3	87	12549	28	1,271	932-1,734	3,129	2,790-3,592
7	S1*S3 S1*S2*S3	89	12548	30	1,718	1,472-2,005	3,576	3,330-3863
8	S2*S3 S1*S2*S3	60	12563	0.01	1,892	1,629-2,197	3,750	3,487-4,055
9	S1*S2 S1*S3	92	12547	32	600	408-881	2,458	2,266-2,739
10	S1*S2 S2*S3	60	12562	0.6	1,866	1,298-2,681	3,724	3,156-4,539
11	S1*S3 S2*S3	60	12563	0.3	1,896	1,615-2,227	3,754	3,473-4,085
12	S1*S2 S1*S3 S2*S3	62	12563	-	2,220	1,245-3,957	4,078	3,103-5,815
13	S1*S2 S1*S3 S1*S2*S3	62	12563	-	600	408-881	2,458	2,266-2,739
14	S1*S2 S2*S3 S1*S2*S3	62	12563	-	1,866	1,298-2,681	3,724	3,156-4,539
15	S1*S3 S2*S3 S1*S2*S3	62	12563	-	1,896	1,615-2,227	3,754	3,473-4,085
16	S1*S2 S1*S3 S2*S3 S1*S2*S3	62	12563	-	2,220	1,245-3,957	4,078	3,103-5,815
17	S1*S2 S1*S3 S2*S3 S1*S2*S3	62	12563	-	2,220	1,245-3,957	4,078	3,103-5,815

[1]Clayton, Cobb, DeKalb, Fulton, Gwinnett

[2]S1=Sibley adolescents, 11-20 years; S2=Combined CHOA/PCS adolescents, 11-20 years; and S3=Medicaid adolescents, 11-20 years

[3]All models build off this original model containing S1, S2, and S3

[4]Predicted missing were added to the total captured in the four data sources (n=1,858)

Table 7. CR Analysis Using Poisson Modeling of Four Adult Sources to Estimate Missing and Total CHD Cases in Five Metropolitan Atlanta, Georgia Counties[1]

Model	Interaction Terms[2]	AIC Criterion	G^2	Deviance / df	Pred. Missing	95%CI Missing	Est. Total[4]	95%CI Total
1[3]	none	386	20882	35	6,406	5,665-7,244	9,589	8,848-10,427
2	S1*S2	287	20933	27	4,830	4,231-5,515	8,013	7,414-8,698
3	S1*S3	319	20917	31	4,828	4,209-5,537	8,011	7,392-8,720
4	S1*S4	265	20944	24	12,969	10,690-15,732	16,152	13,873-18,915
5	S2*S3	360	20896	36	6,163	5,449-6,970	9,346	8,632-10,153
6	S2*S4	382	20885	38	6,236	5,506-7,063	9,419	8,689-10,246
7	S3*S4	317	20918	30	7,499	6,567-8,562	10,682	9,750-11,745
8	S1*S2 S1*S3	169	20992	13	3,096	2,663-3,599	6,279	5,846-6,782
9	S1*S2 S1*S4	232	20961	22	8,903	7,149-11,086	12,086	10,332-14,269
10	S1*S2 S2*S3	255	20950	26	4,585	4,014-5,237	7,768	7,197-8,420
11	S1*S2 S2*S4	288	20933	30	4,795	4,198-5,478	7,978	7,381-8,661
12	S1*S2 S3*S4	237	20959	23	5,630	4,880-6,494	8,813	8,063-9,677
13	S2*S3 S2*S4	357	20899	40	6,009	5,304-6,808	9,192	8,487-9,991
14	S2*S3 S3*S4	294	20930	31	7,192	6,298-8,213	10,375	9,481-11,396
15	S1*S2 S1*S3 S1*S4	171	20993	15	3,299	2,502-4,349	6,482	5,685-7,532
16	S1*S2 S2*S3 S2*S4	257	20950	30	4,576	4,006-5,228	7,759	7,189-8,411
17	S1*S3 S2*S4 S2*S3	279	20939	34	4,375	3,803-5,032	7,558	6,986-8,215
18	S1*S2 S2*S3 S3*S4	208	20974	22	5,319	4,609-6,138	8,502	7,792-9,321
19	S1*S2 S2*S4 S3*S4	236	20960	26	5,571	4,829-6,428	8,754	8,012-9,611

[1]Clayton, Cobb, DeKalb, Fulton, Gwinnett

[2]Adults, 21-64: S1=Emory; S2=St. Joseph; S3=Grady; and S4=Medicaid

[3]All models build off this original model containing S1, S2, S3, and S4

[4]Predicted missing were added to the total captured in the four data sources (N=3,183)

CHAPTER III: Public Health Implications

This study provides a more detailed picture of the total number of CHD cases in five metropolitan Atlanta, Georgia counties between 2008 and 2010. Despite the close approximation of the adolescent CHD cases using CR methodology, it appears that adult CHD cases are more likely to be underestimated when clinical/billing and administrative records are used for surveillance.

When examining captures and recaptures across data sources, it was important to determine the number of overlapping cases between sources. As expected Medicaid had overlap with the other databases; however, the multitude of overlapping cases between the other institutions showed lack of patient retention especially among adolescents. This overlap could cause differences in care due to lack of familiarity with patient history, coordinated care, and effective communication and education efforts by the primary care physician.[56] In the adolescent population, prolonging pediatric care provides opportunity for patient education, patient maturity, and coordinated transfer to adult care.[50]

The proportion of adult CHD cases in the adolescent specific sites provides context that there is a need for more specialized adult congenital cardiac programs in the state and that there is a lack of referral from adolescent care to adult institutions. It has been estimated that 6.3% of pediatric hospital admissions account for roughly $1 billion a year hospital charges related to adult care.[51] Inadequate spending and utilization of resources coupled with a lack of experience and expertise in the provision of care in congenital heart care by adolescent hospitals to adults who may have comorbidities, decreases the safety of the patient.[51] Informal referrals, long-standing practices, provider attachment, and insurance issues are barriers to adult congenital care.[51] One strategy

favored by pediatricians is taking a multidisciplinary team approach with physicians, nurses, and social workers all educated to meet the adult congenital heart disease patient, including management of comorbidities.[57] Setting strict age policies for transferring patients, addressing the additional staff needed for the labor intensive care required of adult CHD patients, and the creation of referral relationship with a nearby institution would significantly help adult patients transition from pediatrician care to adult coordinated care.

In the current investigation, the majority of missing CHD cases were revealed in the adult age group. Advances in diagnosing and treating children with congenital heart disease have progressed survival into adulthood and clearly this growing population necessitates improved seamless transition of healthcare into adulthood. The literature states that the first gap in care is likely around the age of 19-20 years where patients may be relocating or changing insurance providers, and that that gap is around three years.[5] However, preliminary data from the larger study from which this investigation was based suggested that the gap occurs earlier in adolescence, around 16 years old. Patients with gaps in care at this age have more need for urgent cardiac interventions as seen in the increase in emergency room visits for this age group or have under treated cardiac-related conditions.[5] While the American College of Cardiology/Adult Congenital Heart Disease Association (ACC/ACHD) guidelines state that planned processes addressing the education, medical, psychosocial, and vocational needs of young adults required as they transition from child-centered to adult-oriented care should begin at the age of 12, a large proportion of the providers are not aware of these resources or do not sufficiently meet these guidelines.[5,56,57] In order to reduce the number of emergency room visits by

adolescents and prevent gaps in care, providers should be aware of the four PATCH challenges PATCH (Provider Action for Treating Congenital Hearts). These include increased awareness of ACC guidelines, networking between ACHD specialists and general internists, access to ACHD centers for excellence, and increased educational resources for this population.[58] Gaps in care caused by discrepancies in physician suggestions for follow-up, guidelines for treating complex or simple CHD cases, geographic barriers, and lack of symptoms can all be acknowledged through enhanced educational resources and additional training programs for healthcare workers that treat this population.[5]

Despite the need for lifelong care, substantial gaps in care often occur. For instance, many adolescents with CHD are either lost to follow-up or are not receiving the recommended care as they transition to adulthood.[1] Between 2002-2005, a population of adults who were seen in an Adult Congenital Heart Disease regional clinic had gaps in care ranging from 2 to 50 years with a median lapse of medical care of 10 years.[3] The gap was defined as time from leaving pediatric cardiac care to accessing subsequent cardiac care. It is important that continuity of care be established for optimal quality of life, improvement of health outcomes, and improvement of medical efficiency. Practice variation, lack of patient/parent awareness, and provider availability are some of the barriers to transitioning into adult cardiac care.[1] Barriers limiting lifelong care are likely multifactorial in nature making strategies for identifying possible solutions more complex to both implement and study.

On March 23, 2010, the Affordable Care Act was signed into law, making health care coverage more available, affordable, and adequate for patients with heart disease and

stroke.[59] Those with a pre-existing condition, like a congenital heart disease, can no longer be denied healthcare coverage and are able to remain on their parents' health insurance until the age of 26. The law also includes the provisions of the Congenital Heart Futures Act, which will improve the nation's surveillance, research, and education efforts to fight congenital heart disease.[59] During the time this study was conducted, healthcare reform policies had not yet been put in place. Since this expansion was not yet in place, prevalence estimates post March 2010 could largely change with establishment of the Affordable Care Act.

Understanding and implementing optimal health services in a systematic manner provides an opportunity to improve health outcomes for patients with CHDs. Ensuring access to age appropriate care, facilitating the transition of care from child to adult oriented care, and improving the quality of care by minimizing wasted expenditures will improve the delivery of health services and care to persons with CHD.[1]

APPENDICES

Appendix A: Congenital Heart Defects Case Definition

For an adolescent or adult with a CHD to be included, the following criteria must be met: must have at least one of the following CHD ICD-9 codes within 745-747, 648.5, 648.6, V42.1, 996.83; must have been seen in at least one of the eight healthcare facilities from which we are receiving data between 2008-2010; must be at least 11 years of age as of 1/1/2010; and must live in the state of Georgia.

Birth Defects	**ICD-9-CM Codes**
Pregnancy associated with cardiac conditions	648.5
Pregnancy associated with cardiac conditions	648.6
Bulbus cordis anomalies & anomalies of cardiac septal closure	745
Compl transposition of great vessels	745.10
Double outlet right ventricle, Dextratransposition aorta, Incomp	745.11
Corrected transposit great vessels	745.12
Transposition great vessels; other	745.19
Tetralogy of Fallot, Fallot's pentalogy	745.22
Common ventricle, Cor triloculare biatriatum, Single ventricle	745.3
Ventricular septal defect, Left ventricular-right atrial communic	745.43
Ostium secundum type atrial septal defect, Defect: atrium secundum	745.54
Atrioventricular septal defect (endocardial cushion defect)	745.6
	746.61
Endocardial cushion defects; other	745.69
Cor biloculare, Absence of atrial and ventricular septa	745.7
Bulbus cordis anomalies & cardiac septal closure; other	745.8
Other congenital anomalies heart; Pulmonary valve anomaly, unspec	746
Atresia, congenital, Congenital absence of pulmonary valve	746.01
Stenosis, congenital	746.02

Anomal pulmon valve; othr, Congen insufficiency pulmon valve, Fallot's	746.09
Tricuspid valve atresia & stenosis	746.15
Ebstein's anomaly	746.2
Congenital stenosis of aortic valve, Congenital aortic stenosis	746.3
Congenital insufficiency of aortic valve, Bicuspid aortic valve, Congenital aortic insufficiency	746.4
Congen mitral stenosis, Fused commissure mitral valve, Parachute deform mitral valve, Supernum cusps	746.5
Congenital mitral insufficiency	746.6
Hypoplastic left heart syndrome, Atresia, or hypoplasia aortic orifice/valve, hypoplasia ascend aorta & defective develop left ventricle (w mitral valve atresia)	746.75
Other specified anomalies of heart	746.85
Subaortic stenosis	746.81
Cor triatriatum	746.82
Infundibular pulmonic stenosis, Subvalvular pulmonic stenosis	746.83
Obstructive anomalies heart, NEC, Uhl's disease	746.84
Coronary artery anomal, Anomalous origin/commun coronary artery, Arteriovenous malform coronary artery: absence, aorta or pulmon	746.85
Congen heart block, Compl or incompl atrioventri [AV] block	746.86
Malposition of heart and cardiac apex, Abdominal heart, Dextrocardia, Ectopia cordis, Levocardia (isolated), Mesocardia,	746.87
Spec anomal heart; other, Atresia cardiac vein, Hypoplasia cardiac vein, Congen: cardiomegaly, divert, left ventr, pericardial defect	746.895
Unspec anomaly heart, Congen: anomaly heart NOS, heart disease NOS	746.9
Other congen anomalies circ sys	747
Patent ductus arteriosus, Patent ductus Botalli, Persist ductus arteriosus	747
Coarctation of aorta	747.1

Coarct of aorta (preductal) (postduct), Hypoplasia aortic arch	747.106
Interruption of aortic arch	747.11
Other anomalies of aorta	747.2
Anomaly of aorta, unspecified	747.2
Anomaly aortic arch, Anomal orig	747.21
Atresia & stenosis aorta, Absence or Aplasia aorta	747.22
Anomalies aorta; other, Aneurysm sinus Valsalva	747.29
Anomalies of pulmonary artery	747.3
Pulmonary artery coarct & atresia	747.31
Pulmonary arteriovenous malform	747.32
Other anomal pulmon artery & pulmon circ	747.39
Anomalies of great veins	747.4
Anomaly great veins, unspec, Anomaly NOS pulmon veins, vena cava	747.4
Total anomalous pulmon venous connection, Total anomalous pulmonvenous return [TAPVR]: subdiaphragm, supradiaphragm	747.41
Partial anomal pulmon venous connection, Part anomal pulmon venous return	747.42
Other anomalies great veins, Absence vena cava (inferior) (superior), Congen stenosis vena cava (inferior/superior), Persist: left post cardinal vein, left super	747.49
Absence/hypoplasia umbilical artery, Single umbilical artery	747.5
Other anomalies of peripheral vascular system	747.6
Other spec anomalies circulatory sys	747.8
Anomalies cerebrovascular sys, Arteriovenous malformation brain	747.81
Spinal vessel anomaly, Arteriovenous malform spinal vessel	747.82
Persistent fetal circ, Persistent pulmon hyperten, Primary pulmon hyperten newborn	747.83
Specified anomalies circ sys; other, Aneurysm, congen, spec site not elsewhere classified	747.89

Unspec anomaly circulatory sys	747.9
Heart transplant codes	V 42.1
Heart transplant codes	996.83

Appendix B: Tabulated Literature Review

Table1. Papers Explaining Capture-Recapture Methodology

Study	Approach	Major Points
Chao, et al., 2001	Discusses three classes of capture recapture: ecological modeling (best when >3 trappings), log-linear modeling, and sample coverage approach.	Discusses use of CARE program for analysis which requires S+. Also discusses the instability of estimates when lack of overlap between sources. Use of log-linear modeling provides uniform framework and implementation can be easily applied.
Hook and Regal, 1997	Log-linear modeling the model selected has major implications with the estimate. Variations in model uncertainty result from use of different criterion (AIC, BIC, Draper, Schwarz).	The saturated model often appears optimal and investigators should use caution if selecting this model especially in circumstances of sparse cells. AIC criteria outperformed the other criteria in the simulations.
Hook and Regal, 1999	Offer general recommendations of presentation and approach to method for epidemiologists. The aim and use of the estimate, along with the target population should be kept in mind during the entire process.	Identify sources and their characteristics, paying close attention to possible relationships between sources. Examine the structure of the data and before more complex approaches cover two way capture recapture and describe the data. When undertaking log-linear modeling pertinent covariates should be considered carefully.
Alho, 1990	Heterogeneity in capture probabilities can bias results and create incorrect estimates when using two-source capture recapture approach. This paper recommends use of logistic regression modeling to overcome this bias and offers a standard notation to try and create a familiar representation of the method.	A long statistical derivation of the variance and regression of the estimator are undertaken. Using simulations, the authors show that the bias of the estimated total sample is less then when using the classical analysis; however, the variance of the estimate is larger to overcome this bias.
Wittes, 1974	Standard notation and definition of the random variables are proposed in this paper. Small cell corrections are discussed along with measures of variance. Further investigation of list bias is discussed.	When the assumption of independence questioned, equations can be used to determine if independence is not met. Equations using the probability of inclusion on a list can be done to check.

Table 2. Papers Utilizing Capture-Recapture Methodology for Health Assessment

Study	Health Outcome	Sample	Sources	Linkage	Capture-Recapture	Modeling
Wittie, M, et al (2013)	ALS	ALS patients living in 5 county area of metropolitan Atlanta	4 sources: Clinical sources (Emory Healthcare, etc.), VA (VHA and VBA), ALSA membership, and Georgia mortality records	Yes, used probabilistic algorithms (last name, first name, SSN, DOB)	Two-source CR conducted to assess dependency between sources	Log-linear modeling performed under Poisson, stratified by age and race (W, NW)
Bruno, et al. (1994)	Diabetes (IDDM and NIDDM)	Residents of Casale Monferrato with diagnosis on October 1, 1988	4 sources: diabetes clinic, discharge hospital data, computerized database of prescriptions, list of reimbursement data	Not specified	Two-source CR conducted to assess dependency between sources	Log-linear models with stratificatio n by age-group and pattern of treatment
Corrao, et al.(2000)	Alcohol related disorders	Residents of Voghera aged >15 years (according 1991 census) received treatment throughout 1997	4 sources: self-help volunteering groups, psychiatric ambulatory, public alcohology services, and computer database of patients discharge records from hospital	Yes, used 20 digit code (initials of surname and name (2), gender (1), DOB (6), municipality of birth (5), residence (3), and family physican (5))	Two source CR was conducted to assess total population and dependency between sources	Log-linear models were conducted with stratificatio n assessed by age and gender
Somers, C., et al (2014)	Systemic Lupus Erythema tosus	Residents of 2 counties in Michigan and 1 outside due to pilot data diagnosed during surveillance, 01/01/02-12/31/04.	4 sources: hospitals, rheumatology, nephrology/der matology, ESRD-US Renal system database (Medicaid and laboratory data was dropped)	Not specified	Two-way CR conducted to evaluate number of missing cases from the sources	Log-linear modeling was conducted assuming no 3-way interaction terms

La Ruche, G., et al (2013)	Dengue Fever	Residents of metropolitan France with diagnosed dengue between 2007-2010	3 source surveillance network: laboratory notification network, mandatory physician notification (physicians and biologists), enhanced surveillance system of clinically suspected	Yes, created unique ID (patients DOB, sex, postal code of residence/laboratory collected, date of blood sampling)	Two-way CR conducted to evaluate dependency between sources	Found that the enhanced surveill. was highly dependent on the other two sources so two-source CR was conducted with Choa's estimator and stratified by geographic area and period of the year

Appendix C. Example SAS Code of Capture Recapture Using Poisson Modeling

```
data x;

input source1 source2 source3 count;

cards;

1 1 1 40
1 1 0 288
1 0 1 52
0 1 1 46
1 0 0 1386
0 1 0 394
0 0 1 70
0 0 0 .
;
run;

proc sort data=x;

by source1 source2 source3;

run;

/*model 1 young*/
proc genmod data=x;

model count = source1 source2 source3/dist=poisson link=log obstats lrci;

run;
```

Appendix D. Sensitivity Analysis using MACDP

A separate sensitivity analysis was conducted to investigate the performance of the MACDP (Metropolitan Atlanta Congenital Defects Program) birth registry for inclusion of CHD cases found in MACDP as compared to the other sites. The MACDP has been called the most comprehensive system in the US and has been used as the "gold standard" in comparison to other registries.[38] In 1999, the sensitivity of the MACDP one year after birth and using birth certificate data was found to be 87%.[38] While this estimate was almost a decade ago and included all birth defects, the case finding mechanisms of the program do not collect all cases within this area.

The MACDP uses active-case finding mechanisms to maintain its population birth defects surveillance database and from 1967 to 2012, the metropolitan Atlanta counties included Clayton, Cobb, DeKalb, Fulton, and Gwinnett. For the purposes of this study, cases found in the registry were between the ages of 11-42 years. Due to the county and age limitations, for this analysis the adolescent sources of CHOA, Sibley, PCS, Medicaid, and MACDP were constrained to CHD cases 11-20 years. The adult sources to be consistent with MACDP were constrained to the five counties and CHD cases ages 21-42 years.

Overall, 20% of those adolescents found in MACDP were later identified in one of the adolescent sites as compared to 6% of the adults (Table D1). Out of all the data sites, Sibley had the highest overlap with MACDP (20%) followed by CHOA with 14%. Sensitivities of the MACDP with adult sources was low, ranging from 1-7% when comparing individual sites (Table D1).

Table D1. Breakdown of Frequencies of Overlap between Adolescents in the MACDP Registry and Adolescent Data Sources and Adults in the MACDP Registry and Adult Data Sources

Adolescent Data Sources (11-20 years)	**N**	**Overlap b/w Source and MACDP**	**Combined Total N Source and MACDP**	**% Found in MACDP**
MACDP	364			
CHOA	394	107	758	14%
Sibley	1,384	346	1,748	20%
PCS	10	2	374	1%
Medicaid	70	31	434	7%
Total[1]	1,858			20%
Adult Data Sources (21-42 years)	**N**	**Overlap b/w Source and MACDP**	**Combined Total N Source and MACDP**	**% Found in MACDP**
MACDP	93			
Emory	1,039	65	1,132	6%
Grady	142	3	235	1%
St. Joseph	85	1	178	1%
Medicaid	182	19	275	7%
Total[1]	1,448			6%

[1]Totals excluded MACDP in both the adolescent and adult sensitivity analyses.