DUKECATHR dataset

A dataset of cardiac catheterization procedures created for educational use

Documentation for users

Version: November 10, 2016

Introduction

The Duke Databank for Cardiovascular Disease (DDCD) is a clinical care database created through the vision of Dr. Eugene Stead, chair of the Duke Department of Medicine from 1946 to 1967[1]. The DDCD was established in the 1960's by a research team within the Duke Division of Cardiology, Department of Medicine. The Department received a Myocardial Infarction Research Unit grant to place a computer on the cardiac care unit. In addition to the primary goal of collecting and monitoring data on cardiac care unit patients, data were collected on patients undergoing cardiac catheterization for suspected coronary artery or valvular heart disease and cardiac surgery with the focus on both generating reports for the patients' medical records and to have these data available for clinical research in order to improve the medical care of patients with chronic disease, specifically coronary artery disease. In-hospital data including treadmill, EKG, chest x-ray, nuclear imaging, and Holter monitor results were collected on patients who underwent coronary angiography and coronary artery bypass graft surgery and combined with outcome data collected through long-term follow-up. The original database system (CIRCE/MIDAS) was converted and expanded into The Medical Record (TMR) database in 1984 and 1985 to include Catheterization Lab, Coronary Artery Bypass Graft, and Cardiac Diagnostic Unit noninvasive imaging. Conversion to a relational database known as DISCC (Duke Information System for Cardiovascular Care) occurred in the early 1990's. The Duke Databank and resulting data collection along with the expanding focus on performing randomized clinical trials led to the creation of the Duke Clinical Research Institute (DCRI) in 1996.

[1] Improving patient care by capturing computerized data: a glimpse into the creation of the Duke Databank for Cardiovascular Disease, http://digitaldukemed.mc.duke.edu/databank/

The DDCD has been the basis of many methodological[2] and clinical[3] research studies. To support these efforts, information that is frequently used in research studies has been extracted from the DDCD and stored in SAS® analysis datasets. One such dataset, DUKECATH, is a de-identified file including records and variables for patients undergoing cardiac catheterization procedures. The DUKECATH dataset has been made available to researchers through DCRI's SOAR (Supporting Open Access for Researchers) initiative[4]. Requests to access the DUKECATH dataset can be made through http://www.dcri.org/our-approach/data-sharing/.

The DUKECATHR dataset is a subset of DUKECATH that has been created for educational purposes only. This subset includes selected variables and records from DUKECATH and has been anonymized in order to remove individually identifiable protected health information. The records included in DUKECATHR were intentionally selected in a fashion such that meaningful or publishable clinical interpretations of analyses are not possible. However the dataset maintains features of a real-life dataset such as missing records and unusual values that make it interesting for educational and training purposes. The Duke Medicine Institutional Review Board (Pro00068333 and Pro00076907) approved the creation, anonymization, and distribution of the DUKECATHR dataset to approved educators. Characteristics of the DUKECATHR dataset are described in the following sections of this document.

Cohort

The DUKECATHR dataset includes a subset of cardiac catheterization procedures conducted in adult patients (age ≥ 18 years) at Duke University Medical Center between January 1, 1985 and December 31, 2013. Patients with evidence of significant coronary artery disease diagnosed at catheterization were identified, and non-uniform random sample of these patients was selected for inclusion in DUKECATHR. For each of the selected patients, the first catheterization procedure with evidence of

[2] Evaluating the yield of medical tests (Harrell et. al., JAMA 1982); Investigation of test statistics used with the Cox and logistic models (Lee et. al., Biometrics 1983); Regression modeling strategies for improved prognostic prediction (Harrell et. al., Statistics in Medicine 1984); Predicting outcome in coronary disease: Statistical models vs. expert clinicians (Lee et. al., Am J Med 1986); Using observational data to estimate prognosis: an example using a coronary artery disease registry (DeLong et. al., Statistics in Medicine 2001).

[3] Outcomes in medically treated coronary disease (Harris et. al., Circulation 1979, 1980); Estimating the likelihood of significant and severe CAD (Pryor et. al., Am J Med 1983, 1991); The evolution of medical and CABG therapy for CAD (Califf et. al., JAMA 1989); Comparison of outcomes with medical, CABG and PCI (Mark et al Circulation 1994).

[4] Pencina MJ, Louzao DM, McCourt BJ, Adams MR, Tayyabkhan RH, Ronco P, Peterson ED. Supporting open access to clinical trial data for researchers: The Duke Clinical Research Institute–Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. American Heart Journal. 2016 Feb 29;172:64-9.

significant coronary artery disease was included as well as all subsequent cardiac catheterization procedures.

The dataset contains one record per catheterization procedure. Individual patients who underwent more than one procedure during the study period have multiple records in the dataset: one record for each procedure. Left heart catheterization procedures (typically used to diagnose and treat blockages in coronary arteries) and right heart catheterization procedures (typically used to determine pressures within the heart and lungs) are included. Some procedures involve both left and right heart catheterization. The purpose of a cardiac catheterization procedure may be diagnostic or interventional, and a single procedure may involve both diagnostic and interventional components. For example, left heart catheterization may be used to visualize the extent of blockages in the coronary arteries (diagnostic) or to treat a blockage through percutaneous coronary intervention (interventional). Right heart catheterization may be used to assess heart failure, congenital heart disease, valvular heart disease, cardiomyopathy, or pulmonary hypertension, or to collect a specimen for biopsy, and may involve intravenous interventions with medication during the procedure so that physicians can monitor effect on function. The focus of the DUKECATHR dataset is on coronary artery disease and left heart catheterization results. Characteristics of the procedures included in the DUKECATHR dataset are summarized in Table 1.

Table 1 – Characteristics of cardiac catheterization procedures included in DUKECATHR

Number of catheterization procedures	83,320	
Number of unique patients	39,098	
Catheterization approach	Unknown:	1,711 (2.1%)
	Right heart only:	1,424 (1.7%)
	Left heart only:	71,105 (85.3%)
	Right and left heart:	9,080 (10.9%)
Procedures with arteriograms available	79,643 (95.6%)	
Procedures with percutaneous coronary interventional component	28,091 (33.7%)	

Evaluations performed prior to, or at time of, catheterization procedure

The DUKECATHR dataset includes variables describing the following patient and procedure characteristics evaluated prior to or at the time of cardiac catheterization:

- Year when cardiac catheterization procedure was performed
- Patient demographics: age, race, gender
- Clinical history: cardiovascular risk factors, comorbidities, and days since most recent prior cardiac procedure or event (CABG, PCI, myocardial infarction).
- Vital signs and physical exam findings prior to cardiac catheterization
- Laboratory results (creatinine, LDL cholesterol, HDL cholesterol, Total cholesterol) based on the most recent data available within 1 year prior to cardiac catheterization
- Catheterization results: evaluation of stenosis in major coronary arterial systems and bypass grafts, left ventricular ejection fraction
- Whether coronary intervention was performed during catheterization

Some evaluations were not collected prior to conversion of the DDCD to the DISCC system, and are set to missing in the pre-DISCC era (pre-1994).

Follow-up evaluations

The DUKECATHR dataset includes follow-up for the following cardiovascular events and procedures that occurred through December 2014[5]:

- Death (all-cause)
- Non-fatal myocardial infarction (MI)
- Non-fatal stroke
- Coronary artery bypass graft (CABG) surgery
- Percutaneous coronary intervention (PCI)

[5] In order to preserve anonymization of records in the DUKECATHR dataset, follow-up was censored at an arbitrary date in December 2014.

Only the first follow-up event or procedure that occurs on the day of, or subsequent to, the cardiac catheterization procedure is reported in DUKECATHR. Repeat events are not included. The number of days from the catheterization procedure to the first subsequent event is reported.

Follow-up information on these events and procedures was obtained from Duke University Health System records. In addition, a subset of patients in DUKECATHR were enrolled in an active follow-up protocol. This follow-up cohort included all patients diagnosed with significant coronary artery disease during cardiac catheterization, as well as patients who underwent PCI or CABG surgery at Duke. These patients were contacted by mail and/or telephone at 6 months and at 1 year subsequent to their cardiac catheterization procedure, and annually thereafter. Vital status as well as occurrence of cardiovascular events and procedures that occurred outside the Duke University Health System were ascertained. Vital status for patients who were on active follow-up but who could not be contacted, was queried through the National Death Index (NDI) on an annual basis. From time-to-time, NDI queries for vital status were obtained for patients without significant coronary disease who were included in certain studies, and this information is included in DUKECATHR. The extent of follow-up for vital status is summarized in Table 2.

Table 2 – Extent of follow-up for vital status in DUKECATHR

Patients enrolled in active follow-up protocol		37,747 / 39,098 (96.5%)
Vital status known at 30 days post cath (all cath procedures)		83,230 / 83,320 (99.9%)
Vital status known at 1 year post cath (all cath procedures)		83,082 / 83,320 (99.7%)
Years to death or last contact after first cath		
	n	39,098
	Mean (SD)	9.8 (7.3)
	Median (Q1, Q3)	8.7 (3.6, 14.8)
	Min, Max	0.0, 29.9
Years to last contact after first cath (in patients alive at last contact)		
	n	14,576
	Mean (SD)	12.4 (7.6)
	Median (Q1, Q3)	11.6 (6.1, 17.9)
	Min, Max	0.0, 29.9

Tips for using the DUKECATHR dataset

The accompanying DUKECATHR Dataset Dictionary provides detailed information about variables included in DUKECATHR.

Record identifiers and sequencing: Unique patients are identified by the RSUBJID (Subject ID) variable. For each patient, unique catheterization procedures are identified by the RSEQCATHNUM (Patient's Sequential Cath Number) variable. The number of days that elapsed between the first catheterization procedure reported in DUKECATHR for a particular patient and each subsequent procedure is recorded in the RDAYSFROMINDEX (Days from Index Cath) variable. A patient may have more than one catheterization procedure on the same day, including on the day of the index procedure; consequently multiple procedures may have the same value of RDAYSFROMINDEX. A small number of patients have a large number of catheterization procedures reported in DUKECATHR; for example, four patients underwent 40 or more catheterization procedures between 1985 and 2013. This may occur in patients with complicated disease or treatments, for example heart transplant, or coronary artery bypass graft surgery with multiple subsequent percutaneous coronary interventions with stenting.

Identifying patients on follow-up protocol: The variable FUPROTCL (Patient on follow-up protocol) is used to identify patients enrolled in an active follow-up protocol. Other patients may have follow-up events and procedures recorded in DUKECATHR, but these records may not be exhaustive; for example, they may not include events and procedures that occurred outside the Duke University Health System.

DP* and DS* variables: DUKECATHR includes a number of variables named DP* or DS*, such as DPCABG and DSCABG. The DP* variables report the days from the catheterization procedure to the most recent event/procedure prior to the catheterization. A missing value indicates that no evidence was reported of a previous event or procedure. The DS* variables report the days from the catheterization procedure to the next event/procedure subsequent to the catheterization. A missing value indicates that no evidence was reported of a subsequent event or procedure through end of follow-up in December 2014. In DUKECATHR, follow-up was censored at an arbitrary date in December 2014 in order to preserve anonymization of records. If a patient was enrolled in an active follow-up protocol, and the DS* variable is missing, then it is reasonable to assume that there was no occurrence of the event through the date when the patients was last known to be alive (DAYS2LKA). The DSMI and DSSTROKE variables only include days to the first subsequent non-fatal event. Fatal events are included in DEATH. Thus analysis of MI or stroke outcomes using DUKECATHR can only

reasonably be conducted by combining these endpoints into a composite endpoint with mortality, for example 'Death or MI'. The DSPCI variable reports the days from the catheterization procedure to first subsequent PCI, which could be a PCI occurring as part of the same catheterization procedure (that is, the catheterization included an interventional component).

Maximum Coronary Stenosis variables: The extent of stenosis in native coronary vessels is reported in the LMST, LADST, PRXLADST, LCXST, and RCAST variables. Each of these variables is associated with a specific anatomic region of the heart muscle. The extent of stenosis in previous coronary artery grafts is reported in the GRAFTST variable. The stenosis variables are numeric variables taking values between 0 (no stenosis) and 100 (complete stenosis), and report the maximum stenosis across the vessels within each coronary arterial region, or across grafts. The level of stenosis within large vessels (≥2mm diameter) is coded as it was reported by the cathing physicians. Stenosis within small vessels (<2mm diameter) is coded with a value of 7 if stenosis≤50%, and as 13 if stenosis>50%. Coronary dominance (left, right, or balanced) is factored into these calculations. For example, in the case of left dominance the level of stenosis in the RCA is downgraded to that of a small vessel. The maximum across all vessels within the system is then calculated using the coded numeric values. During 2007, the scale used to grade stenosis was changed. Before this date, stenosis was graded as 0 (Normal), 5 (<25%), 25%, 50%, 75%, 95%, and 100%. After this date, stenosis was graded as 0 (Normal), 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 99%, and 100%. An explanation of the relationship between these variables and the NUMDZV (Number of Significantly Diseased Vessels in the three major arterial regions) and CORDOM (Coronary Dominance) variables is provided in the DUKECATHR Dataset Dictionary.

Groupings of Race, Age, and Year of Cardiac Cath: Values of certain variables have been aggregated in order to remove individually identifiable information. Race groups other than 'Caucasian' or 'African American' are aggregated into a single group coded as 'Other'; cases where race was unknown or missing are coded as a missing value (variable RACE_G). Age at time of cardiac catheterization is aggregated into approximately 5 year groupings (variable AGE_G). Year of cardiac catheterization is aggregated into the following groups: '1985-1990', '1991-1994', '1995-1998', '1999-2002', '2003-2006', '2007-2010', and '2011-2013' (variable YRCATH_G). The slightly unequal groupings were chosen in order to retain information about significant changes in care or diagnosis of coronary artery disease, or changes in data collection. For example, coronary stents were approved by the Food and Drug Administration (FDA) in 1994, cardiac catheterization data collection was converted to the DISCC database in 1994, the first drug eluting stent was approved by the FDA in 2003, and the scale used to report stenosis in coronary arteries was changed to deciles in 2007.