AllClinical

SAS Dataset Name: AllClinical

General description:
These datasets contain data from smaller clinical datasets that have one row per participant combined into one dataset for each visit. MIF, MRI, and Xray are not included as these datasets contain multiple rows per participant. Data from the Enrollees dataset are not included because the Enrollees dataset is updated for each data release and would therefore become obsolete for older AllClinical datasets.

Previously, only the smaller functional datasets that grouped data by type were available for download (see list of datasets below). In addition to these smaller datasets, AllClinical datasets are available as of release 2.2.2, 3.2.1, and 4.2.1. These datasets were created to provide an alternative for analysts who want the universe of OAI clinical data. Instead of having to download multiple, smaller datasets, now analysts only need to download the AllClinical dataset, along with the MIF, MRI and Xray datasets to acquire the complete clinical dataset for the visit of interest. The analyst who is interested in a small subset of variables should consider downloading the smaller, functional dataset that contains these variables. For example, if the analyst is interested in only the knee examination variables, the PhysExam dataset for the visit of interest should be downloaded. Although the knee examination variables are also in AllClinical for that visit, the PhysExam dataset is will be more manageable in size with fewer variables than the AllClinical dataset, and will be faster to download. The analyst should always download the most recent version of the Enrollees dataset when working with OAI data.

Detailed information exists for each of the smaller functional datasets whose data comprise the AllClinical datasets.

- Accelerometry
- Biomarkers
- Joint symptoms/function
- Medical history
- Nutrition
- Physical exam, measurements
- Subject characteristics, risk factors

These dataset descriptions apply to all visits and all releases.

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments."
- **Dataset strengths/weaknesses:** AllClinical datasets are very large, as they contain the data from all of the smaller functional datasets that are one-row-per-participant datasets. Analysts interested in a small subset of variables should consider downloading the smaller dataset that contains the variables of interest.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Accelerometry

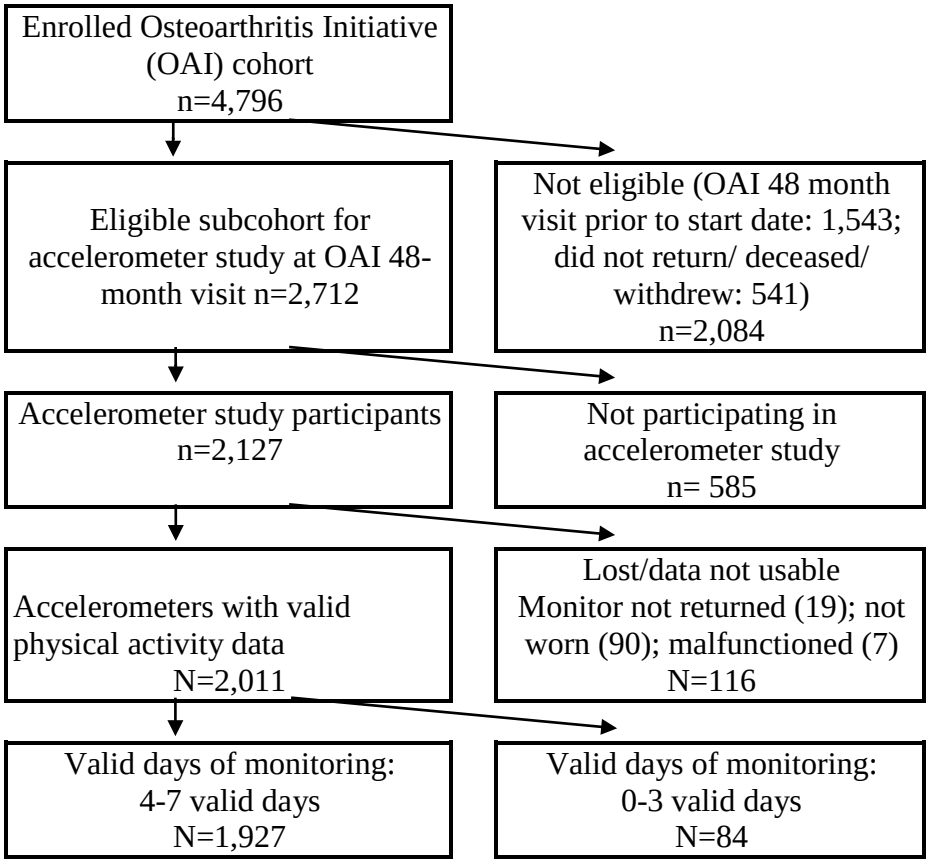
SAS Dataset Name: Accelerometry

General description:
An Osteoarthritis Initiative (OAI) physical activity ancillary study (Dr. Dorothy Dunlop, AS05-10) collected accelerometer data on a subgroup of OAI participants. Physical activity was monitored using ActiGraph GT1M uniaxial accelerometers (ActiGraph; Pensacola, FL). As shown in Figure 1 below, a subset of 2,712 OAI participants were invited to join this ancillary study. The number of OAI participants who consented to participate was 2,712. Of these 2,011 adults returned their monitor with activity data. A valid day of accelerometer monitoring data was determined from recording evidence the monitor was worn at least 10 hours/day. A total of 1,927 participants had 4-7 valid days of monitoring, which is sufficient to estimate typical physical activity (1).

The accelerometry dataset contains for each person the average over the valid monitoring days of each physical activity measure and provides variables that indicate whether or not physical activity guidelines were attained during the week of monitoring.

The detailed minute-by-minute and day-by-day accelerometry data are in separate accelerometry datasets.

Figure 1. Flow chart of OAI participants in OAI 48-month accelerometer study



Accelerometer monitoring in the OAI:
Eligibility for accelerometer monitoring required a scheduled OAI 48-month follow-up visit between August 2008 and July 2010, with staggered starting months across the OAI sites. A total of 2,127 persons consented to participate in accelerometer monitoring representing 78.4% of eligible participants (2,712). Another 1,543 OAI participants had visits that preceded the accelerometer study start date and 541 were deceased/did not return at 48 months/ withdrew from OAI study. Physical activity was objectively measured using a GT1M ActiGraph accelerometer. Trained OAI research staff gave uniform scripted in-person instructions. Each participant was told to wear the accelerometer on a belt at the natural waistline on the right hip in line with the right axilla upon arising in the morning and continuously until retiring at night, except during water activities, for seven consecutive days. Participants maintained a daily log to record time spent in water and cycling activities, which may not be fully captured by accelerometers. Daily log data (not electronically archived) indicated that participants spent in little time in water and cycling activities (median 0 minutes/day, interquartile range = 0.0 to 3.4 minutes/day); this information indicates little activity was missed or underestimated by accelerometer monitoring. Participants returned the accelerometers to the research center; where data were downloaded using the manufacturer's software, and checked for valid data recording.

A total of 2,001 adults provided one or more valid days of accelerometer monitoring data. A valid data of accelerometer monitoring was based on recording evidence that the monitor was worn at least 10 hours.

Accelerometers used in the OAI:
Unlike pedometers, which measure steps, but give no information about the intensity of those steps, accelerometers constantly sample activity for accelerations, and are therefore able to provide information on all three dimensions of physical activity (frequency, intensity, and duration). Physical activity in this OAI sample was monitored in all study participants using a GT1M ActiGraph uniaxial accelerometer. The GT1M ActiGraph is a small uniaxial accelerometer that measures vertical acceleration and deceleration (2). The accelerometer acceleration signal is filtered and digitized by an 8-bit analog-digital (A-D) converter at 30 samples per second. The A-D converter measures the magnitudes of the captured accelerations.

The output from an accelerometer is an ***activity count*** (explained below). Spurious accelerometer counts were identified by negative counts (<1/1,000,000 recorded negative activity counts); these spurious values were set to missing on a minute by minute basis. Accelerometer accuracy (walking speed(3)) and test-retest reliability(4) of under field conditions have been established in many populations including persons with OA (3-7). Uniaxial accelerometer validation studies against "gold standard" whole-body indirect calorimetry showed high correlation with metabolic equivalent (r=0.93) and total energy expenditure (r=0.93) (8).

What is an activity count?:
An activity count is the weighted sum of the number of vertical movements measured over a time period (e.g. in this case 1 minute), where the weights are proportional to the magnitude of ***measured*** acceleration or deceleration. In contrast to pedometers, an accelerometer measures acceleration and changes in acceleration. Gravity's acceleration is called Gs. For example, when you accelerate your car, the Gs push you back into the seat and vice versa when you hit the brakes (the harder you hit pedals the more noticeable the Gs). Conceptually, accelerometer

counts increase with the increased forces. As you walk, you go up a little and then you come back down, running makes this vertical 'up and down' movement more noticeable. As your movement becomes more noticeable, the counts increase. It is precisely this capability that enables accelerometers to provide data on all three dimensions of physical activity: frequency, duration, and intensity, since data is collected in 'real time'.

Identifying valid days of accelerometer monitoring:
An important analytical step is the translation of accelerometer counts into physical activity measurements. We used methodology validated in adults with osteoarthritis (9). A basic building block in this process is the assessment and interpretation of 'nonwear time'. Nonwear relates to periods when activity counts register as '0', because the accelerometer is not being worn. The challenge is to distinguish '0' due to no activity from '0' due to the monitor not being worn and lying on a table. Analytically, non-wear periods were defined as ≥90 consecutive minutes with zero activity counts (allowing for interruptions of up to 2 consecutive minutes with counts<100)(10). A valid day of accelerometer monitoring was defined as 10 or more wear hours in a 24-hour period (1). Note that the 10 hours of wear time was not required to be continuous.

Accelerometer cutpoints to assess physical activity intensity:
Accelerometer 'cutpoints' are thresholds assigned by researchers to divide activity counts into physical activity intensity levels (e.g., light, moderate, and vigorous). The intensity of physical activity is defined in terms of energy expenditure measured in metabolic equivalent task (MET) units. Cut points are estimated to identify activity counts that corresponds to designated energy expenditure values (e.g., light 1.5 to <3 METS, moderate 3 to <6 METS, vigorous ≥6 METS).

Table 2. Accelerometer cut points for Metabolic Equivalent of Task (MET) energy expenditure values

	Activity Count/Minute Cut Points		
Gross MET Value/Intensity	Troiano(1)	Freedson(11)	Swartz(12)
1.5 to<3 METs (light)	100- 2019	100-1951	100-573
3 to 6 METs (moderate)	2020-5998	1952- 5724	574-4944
≥6 METs (vigorous or greater)	5999+	5725 +	4945+

There are many established/validated cutpoints. We applied cutpoints most commonly cited in the literature, which are shown in Table 2. A benchmark set of cutpoints are those published by Troiano,(1) which were applied to the general adult population from the National Health and Nutrition Examination Study (NHANES). We also provide physical activity outcomes based on other cutpoints due to their importance from earlier publications, which includes Freedson (11) cutpoints for the general adult population and Swartz (12) cutpoints for older adult populations. What distinguishes these thresholds is the age of the participants tested and the physical activity engaged in to arrive at the regression equation that established the cutpoints, e.g. treadmill vs. community activity. The data released for OAI participants have been processed applying the cutpoints published by Swartz (12), Freedson (11), and Troiano (1) to each minute of accelerometer output.

The effect of cutpoint choices can best be demonstrated by graphically applying the values to identical accelerometer output, as shown in Figures 2-4 below. Cutpoints appear as horizontal lines drawn through the data. Notice how activity occurring at identical count levels can be

deemed either ‘light’ (counts occurring above the green line but under the brown line) or ‘moderate’ (counts occurring above the brown line but under the red line) intensity, depending on the cutpoints utilized. Deciding a priori which cutpoints to use in data analysis is a critical decision and will affect your findings. It is therefore important to include a data cutpoint citation in reporting accelerometer data, and a justification for its use in the population of interest.

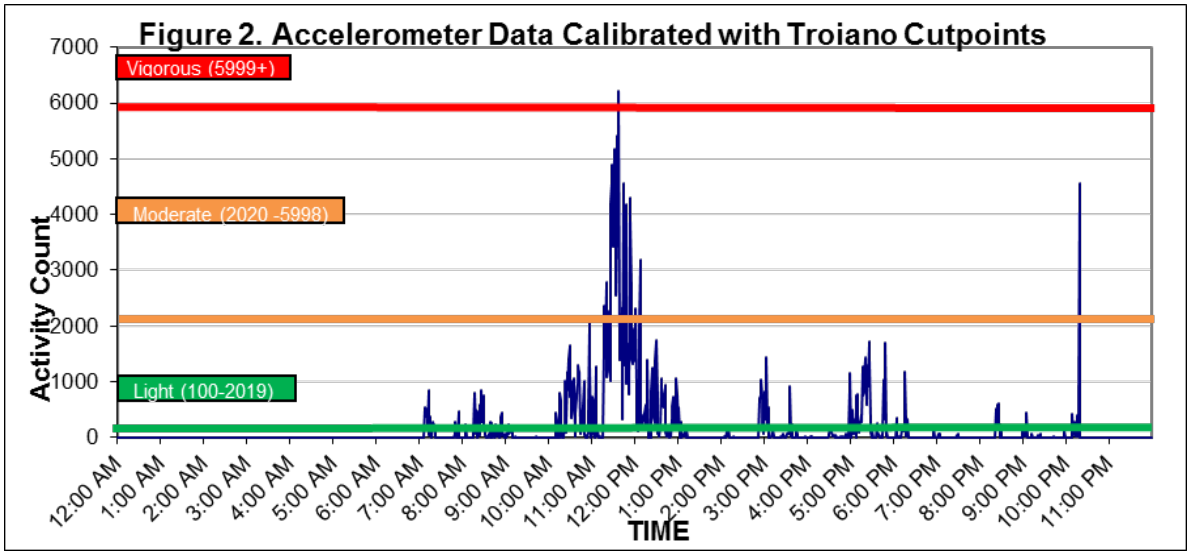


Figure 2. Accelerometer Data Calibrated with Troiano Cutpoints
Vigorous (5999+)
Moderate (2020 -5998)
Light (100-2019)
Activity Count
TIME

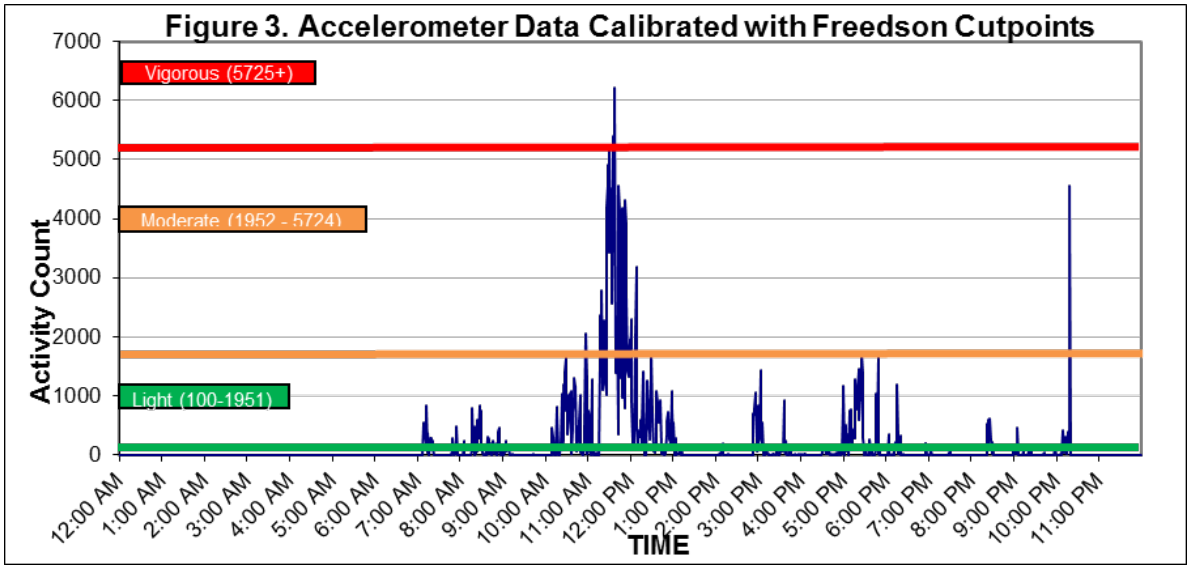


Figure 3. Accelerometer Data Calibrated with Freedson Cutpoints
Vigorous (5725+)
Moderate (1952 - 5724)
Light (100-1951)
Activity Count
TIME

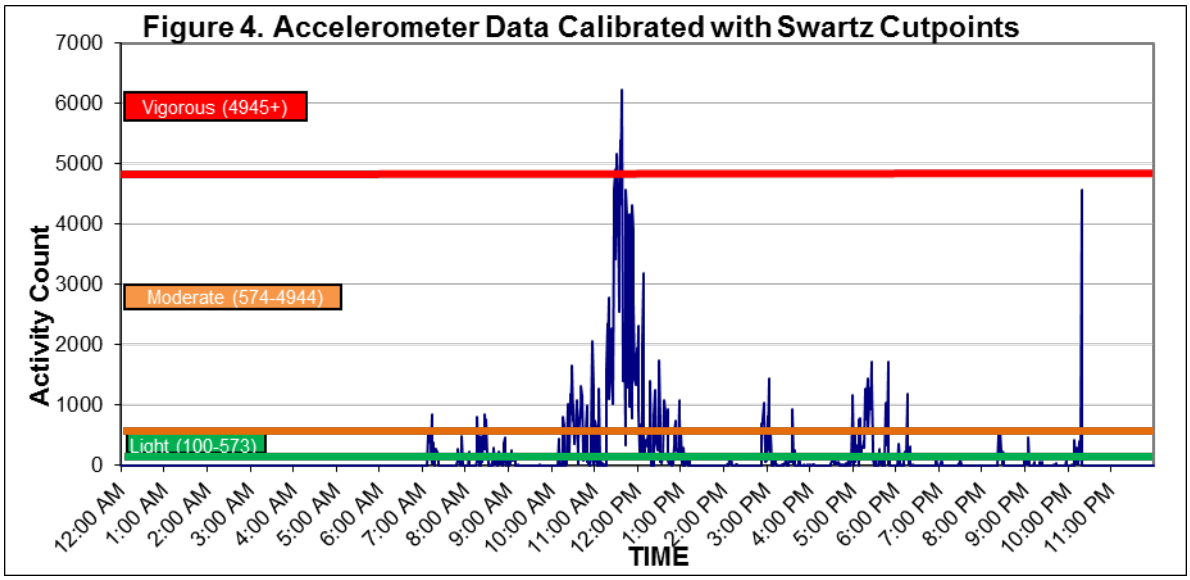


Figure 4. Accelerometer Data Calibrated with Swartz Cutpoints
Vigorous (4945+)
Moderate (574-4944)
Light (100-573)
Activity Count
TIME

Physical activity guidelines:
Physical activity guidelines have been periodically updated since the initial 1995 CDC-ACSM Guidelines (Centers for Disease Control-American College of Sports Medicine) were published in JAMA by Pate et al (13). In 1996 the Surgeon General Guidelines were published (14). In 2003, a committee was convened to update the guidelines, with additional input from the American Heart Association. The final recommendations from the 2003 committee were published in 2007(15). However, the 2008 Physical Activity Guidelines for Americans, is the first official US government policy on physical activity recommendations for optimal health, issued by the U.S. Department of Health and Human Services (DHHS), written by the Physical Activity Guidelines Advisory Committee (PAGAC), based on the scientific report assembled by an independent Federally Appointed Committee of Advisors. The full set of recommendations can be viewed at http://www.health.gov/paguidelines/guidelines/default.aspx#toc. Highlights pertaining to this data release are given in Table 3.

Table 3. Summary of Physical Activity Guidelines 2003-2008

Population Group	**Aerobic Guidelines**		**Strength Training Guidelines***
	Moderate Intensity Activity	Vigorous Intensity Activity	
General Adults (ACSM, 2003)	30 moderate bout** minutes on each of 5 days/week	20 vigorous bout** minutes on each of 3 days/week	2 or more days/week, all muscle groups
General Adults (DHHS, 2008)	150 moderate bout **minutes spread across the week	75 vigorous bout** minutes spread across the week	2 or more days /week, all muscle groups
Adults with Arthritis (DHHS, 2008)	150 moderate-to-vigorous bout** minutes spread across the week***		2 or more days /week, all muscle groups

*The OAI did not collect data to assess attainment of strength training guidelines.
**Minutes are accumulated in 10 minute bouts
***Activities are recommended to be low impact, not painful, and low risk of joint injury.

Aerobic guideline attainment programming decisions:
The attainment of the aerobic guidelines was based on a typical week of physical activity. However, for some participants the number available of valid monitoring days was less than seven days. In those cases, we followed the literature which supports four days as a standard minimum monitoring time needed to capture typical physical activity patterns. Our approach is consistent with the methodology applied to the NHANES data to assess guideline attainment (1). This process is summarized in Table 4.

Table 4. Assessed Guideline Attainment Based on Available Valid Monitoring Days	
Aerobic Guideline	Number of valid days
General Adults (ACSM, 2003)	• For persons with 7 valid days of data (n= 1511, 75.5%) guideline attainment was determined according to tabled outline. • For persons with 4-6 valid days of data (n=317, 20.8%), we estimated the probability that they would attain guidelines by the end of 7 days, using NHANES methodology.(1) • For persons with 0-3 valid days of data, (N=73, 3.7%), no guideline attainment was assessed.
General Adults (DHHS, 2008) Adults with Arthritis (DHHS, 2008)	• For persons with 7 valid days of data (n= 1511, 75.5%) guideline attainment was determined according to tabled outline. • For persons with 4-6 valid days of data (n= 317, 20.8%), we used the average daily physical activity experience to estimate 7 days of activity for the purposes of determining guideline attainment(16) • For persons with 0-3 valid days of data, (n= 73, 3.7%), no guideline attainment was assessed.

Data Structure/Contents:
These datasets contain one record per participant. Each participant included had 4-7 valid days of accelerometer monitoring, i.e. with 10+ wearing hours. The variable uniquely identifying a record is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

The 72-month accelerometry data was released in 2016.

Condition of data:
- **Known data errors:** None at this time. Systematically inspected, cleaned, processed.
- **Dataset strengths/weaknesses:** None at this time.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

References:

1. Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008;40(1):181-8. PubMed PMID: 18091006.
2. Matthews CE, Ainsworth BE, Thompson RW, Bassett DR, Jr. Sources of variance in daily physical activity levels as measured by an accelerometer Med Sci Sports Exerc. 2002;34:1376-81. PubMed PMID: 12165695.
3. Brage S, Wedderkopp N, Franks PW, Andersen LB, Froberg K. Reexamination of validity and reliability of the CSA monitor in walking and running. Med Sci Sports Exerc. 2003;35(8):1447-54. PubMed PMID: 12900703.
4. Welk GJ, Schaben JA, Morrow JR, Jr. Reliability of accelerometry-based activity monitors: a generalizability study. Med Sci Sports Exerc. 2004;36(9):1637-45. PubMed PMID: 15354049.
5. Hendelman D, Miller K, Baggett C, Debold E, Freedson P. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. Med Sci Sports Exerc. 2000;32(9 Suppl):S442-9. PubMed PMID: 10993413.
6. Munneke M, de Jong Z, Zwinderman AH, Tijhuis GJ, Hazes JM, Vliet Vlieland TP. The value of a continuous ambulatory activity monitor to quantify the amount and intensity of daily activity in patients with rheumatoid arthritis 4. J Rheumatol. 2001;28:745-50. PubMed PMID: 11327244.
7. Farr JN, Going SB, Lohman TG, Rankin L, Kasle S, Cornett M, et al. Physical activity levels in patients with early knee osteoarthritis measured by accelerometry. Arthritis Rheum. 2008;59(9):1229-36. PubMed PMID: 18759320; PubMed Central PMCID: PMC2595140.
8. Grenon S, von Specht M, Corso A, Pace J, Regueira M. [Distribution of serotypes and antibiotic susceptibility patterns of Streptococcus pneumoniae strains isolated from children in Misiones, Argentina]. Enferm Infecc Microbiol Clin. 2005;23(1):10-4. Epub 2005/02/11. doi: 13070402 [pii]. PubMed PMID: 15701326.
9. Song J, Semanik P, Sharma L, Chang RW, Hochberg MC, Mysiw WJ, et al. Assessing physical activity in persons with knee osteoarthritis using accelerometers: Data from the Osteoarthritis Initiative. Arthritis Care Res (Hoboken). 2010;62(12):1724-32. Epub 2010/09/02. doi: 10.1002/acr.20305. PubMed PMID: 20806273; PubMed Central PMCID: PMC2995807.
10. Song J, Semanik P, Sharma L, Chang RW, Hochberg MC, Mysiw WJ, et al. Assessing physical activity in persons with knee osteoarthritis using accelerometers: Data in the osteoarthritis initiative. Arthritis Care Res (Hoboken). 2010. Epub 2010/09/02. doi: 10.1002/acr.20305. PubMed PMID: 20806273.
11. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine & Science in Sports & Exercise. 1998;30(5):777-81.
12. Swartz AM, Strath SJ, Bassett DR, Jr., O'Brien WL, King GA, Ainsworth BE. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites Suppl. Med Sci Sports Exerc. 2000;32:S450-6. PubMed PMID: 10993414.
13. Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, et al. Physical activity and public health: A recommendation from the Centers for Disease Control and

Prevention and the American College of Sports Medicine. Journal of the American Medical Association. 1995;273(5):402-7.
14. U.S. Department of Health and Human Services. Physical Activity and Health: A Report of the Surgeon General Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, 1996.
15. Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc. 2007;39(8):1423-34. PubMed PMID: 17762377.
16. Dunlop DD, Song J, Semanik PA, Chang RW, Sharma L, Bathon JM, et al. Objective physical activity measurement in the osteoarthritis initiative: Are guidelines being met? Arthritis Rheum. 2011;63(11):3372-82. Epub 2011/07/28. doi: 10.1002/art.30562. PubMed PMID: 21792835; PubMed Central PMCID: PMC3205278.

Biomarkers

SAS Dataset Name: Biomarkers

General description:
These datasets contain meta-data describing the collection of biospecimens (e.g., which urine void was collected, fasting time before specimen collection, time of day the specimen was collected). If all biospecimens (urine and blood) were not collected on the first try, a repeat collection should have been done. To maintain the one-record-per-participant format, the meta-data variables for the repeat collection were distinguished from those for the first collection by a suffix of 1 (first collection) or 2 (repeat collection). Each collection had a unique barcode, which can be used to identify cryovials filled with specimen from that collection.

The clinical readings of the screening knee x-rays that were done at each clinical site are also included in Biomarkers00. Each BiomarkersXX dataset also contains the variables MRSEQNL and MRSEQNR (the number of unique right and left MR sequences obtained on left and right knees, respectively)

Biospecimen assay data can be found in:

- Biospecimen Assays dataset available in Clinical Data
- Bone Ancillary Study dataset available in both Clinical Data and Image Assessments
- Biospecimens FNIH (LabCorp) dataset available in FNIH
- Biospecimens FNIH (MSBioworks) dataset available in FNIH

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments."
- **Dataset strengths/weaknesses:** None at this time.

General strategies for use:
When using with other datasets, merge/join by ID. Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Joint Symptoms/Function

SAS Dataset Name: JointSx

General description:
The JointSx datasets contain questionnaire results regarding arthritis symptoms in the knee, hip, back, and other joints; arthritis-related joint function and disability; and general health-related function and disability. It also contains data from several standardized instruments for arthritis and general health, including: Western Ontario and McMasters Osteoarthritis Index (WOMAC), the Knee Outcomes of Osteoarthritis Scale (KOOS), and the Medical Outcomes Study (SF12).

Although the category/subcategory organization of the data is independent of the grouping of variables into SAS datasets, the following category groupings of variables are, with only a few exceptions, entirely contained within these datasets:

- Back pain/ all subcategories
- Global function/disability/QOL/ all subcategories
- Knee function/QOL/ all subcategories
- Knee pain/OA status/ all subcategories
- Knee symptoms/ all subcategories
- Other joint symptoms/ all subcategories
- Study eligibility / knee symptoms
- WOMAC/KOOS/ all subcategories

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments"
- **Dataset strengths/weaknesses:** None at this time.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Medical History

SAS Dataset Name: MedHist

General description:
The MedHist datasets contain questionnaire data regarding a participant's arthritis-related and general health histories. Arthritis medical history includes physician diagnoses, joint surgical procedures, joint injuries, screening questions for rheumatoid and inflammatory arthritis, use of conventional and alternative treatments for arthritis, and use of specific medications for arthritis and joint pain. General health history includes comorbidity, depression (CES-D), height and weight history, smoking history, alcohol use, and use of selected non-arthritis medications that may affect joint health and biomarker levels. A complete inventory of a participant's current prescription medications is available in separate datasets (MIF), which contains multiple records per participant (one record per ingredient found in each medication recorded for each participant).

The MedHist datasets contain a number of indicator variables for classes or types of medication (does the participant take that class of medication or not). Some medication use variables are composed of high level terms which encompass multiple medications or a whole class of medications, while some variables refer to specific ingredients. Users interested in a specific indicator variable(s) should review the list of medications/ingredients included in the indicator variables' category. It is possible that new medications have become available since the indicator variables were developed, and are not included in the specific indicator variable(s) of interest. Given this, the MIF datasets should be carefully examined for all prescription medications of interest to the user.

Although the category/subcategory organization of the data is independent of the grouping of variables into SAS datasets, the following category/subcategory groupings are largely or entirely contained within these datasets:

- Anthropometry / Height & weight history
- Joint imaging / MRI exclusions
- Medical history, arthritis/ all sub-categories except family history (see SubjectChar datasets)
- Medical history, general/ all sub-categories (see also current prescription medications in the MIF datasets)
- Medications/ all sub-categories (see also current prescription medications in the MIF datasets)

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record is ID, and the dataset is sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments."
- **Dataset strengths/weaknesses:** For a few variables (WTMAXKG, HT25MM, and PSMKYR), the values for a small number of participants are unusual enough to potentially unmask the identity of the participant. In order to protect these participants, the unusual value has been set to a SAS special missing value code as of release 0.2.1 (".S: Unreleased low value" or ".G: Unreleased high value"). The decision rule for setting these values to missing is given in the release comment for each variable. Use of the SAS special missing values allows these participants to be included in some analyses, either by substituting a uniform small (or large) value, or by dividing the data into quantiles and including .S values in the lowest and .G in the highest quantile.

General strategies for use:

When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Nutrition

SAS Dataset Name: Nutrition

General description:
The Nutrition datasets contains variables collected using the modified Block Food Frequency questionnaire completed by the participant between the screening and the enrollment visits and other nutritional data collected during follow up. As of clinical data release version 0.2.1, the baseline food frequency data have been analyzed to create component nutrient data.

Although the category/subcategory organization of the data is independent of the grouping of variables into SAS datasets, the following category/subcategory groupings are largely or entirely contained within these datasets:

- Nutrition/ all subcategories

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record in these datasets is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments"
- **Dataset strengths/weaknesses:** As of release 0.2.1, the baseline food frequency data have been analyzed to create component nutrient data, also available in the Nutrition00 dataset.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Physical Measurements

SAS Dataset Name: PhysExam

General description:
The PhysExam datasets contain physical measurements of participants, including height, weight, BMI, abdominal circumference, blood pressure, isometric strength, knee and hand exams, walking tests, and other performance measures. Calculated or derived variables are also included. For example, height is measured twice, and then if the difference between the two measurements is greater than a maximum allowed value, it is measured twice more. The calculated variable HEIGHT averages the appropriate measurements, and the individual measures are dropped. BMI is also calculated from HEIGHT and WEIGHT for the user's convenience and to ensure uniform values across analyses. In addition, the variable BMICAT is available as of data release version 3.2.1. It categorizes BMI values into five categories (underweight, healthy, overweight, obese, and morbidly obese).

Although the category/subcategory organization of the data is independent of the grouping of variables into SAS datasets, the following category/subcategory groupings are largely or entirely contained within these datasets:

- Anthropometry/ all subcategories except Height & weight history
- Blood pressure & pulse/ all subcategories
- Hand exam/ all subcategories
- Knee exam/ all subcategories
- Performance measures/ all subcategories
- Strength measures/ all subcategories

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record in these datasets is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments."
- **Dataset strengths/weaknesses:** Good Strength isometric chair measurements of upper leg strength are generally taken from electronic data captured by the instrument. However, in some cases, the data were not saved, were garbled by software glitches, or were not identifiable. As a backup, the examiners recorded the most important values on the exam form, and these values are substituted when the electronic data are missing. Values not recorded on the form are then missing (set to

.L) for these individuals. Because there were ongoing software questions as of clinical data release version 0.1.1, only max force production for each of the four muscle groups tested (REMAXFP, RFMAXFP, LEMAXFP, LFMAXFP) were released. Starting with release 0.2.1, the software issues were resolved and six additional performance variables are released for each of the four muscle groups.

After the study started, the Observational Study Monitoring Board suggested tightening the exclusion criteria of the 400-meter walk for safety reasons. Thus, some participants who did the 400-meter walk early in the study might have been excluded if they had been enrolled later in the study. It is difficult to gauge how much this affected exclusions from this exam, but since a total of only 47 participants were excluded based on these criteria (old or new), it is unlikely to have skewed the results significantly.

For variable HEIGHT, the values for a small number of participants are unusual enough to potentially unmask the identity of the participant. In order to protect these participants, the unusual value has been set to a SAS special missing value code as of release 0.2.1 (".S: Unreleased low value" or ".G: Unreleased high value"). The decision rule for setting these values to missing is given in the release comment for each variable. Use of the SAS special missing values allows these participants to be included in some analyses, either by substituting a uniform small (or large) value, or by dividing the data into quantiles and including .S values in the lowest and .G in the highest quantile.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.

Subject Characteristics

SAS Dataset Name: SubjectChar

General description:
The SubjectChar datasets contain demographic information and other descriptive information about enrolled OAI participants collected from questionnaires administered during pre-screening, screening, and clinic visits. Questions pertaining to some of the OA risk factors and other self-report study eligibility questions can also be found here. Medical history information and information about knee symptoms, even if these were factors used in determining study eligibility, are contained in the MedHist and JointSx datasets, respectively. Physical measurements, such as height and weight, are found in PhysExam.

Some variables originally released in SubjectChar00 do not change from visit to visit, but are needed in analyses with data from every visit, so they were moved from this dataset and put into a new non-visit-specific dataset called Enrollees. As of release version 0.2.1, they can be found only in the Enrollees dataset. The variables that moved include race, sex, cohort assignment, site, and image groups. The variable V00IMGSMPL was also renamed to V00IMAGESA when it was moved to Enrollees to make it more consistent with later image-group indicator variables.

The SubjectChar datasets contain the VxxVISITYP variable which provides information on the type of follow-up contact the participant had. At the OAI 108-month and later contacts, if a participant was only willing to complete the questionnaire as a self-administered instrument, the questionnaire was mailed to the participant. Participants that completed the designated contact via the self-administered questionnaire are denoted as VxxVISITYP=5 (Mailed self-administered questionnaire).

At the 96-month Follow-up Visit, the 96-month Close-out Follow-up Interview was administered to those participants who have withdrawn from the study and agreed to the 96-month contact. The interview was mostly administered to participants over the telephone. However, if the participant was only willing to complete the questionnaire as a self-administered instrument, the questionnaire was mailed to the participant. ALL participants who had a 96-month Close-out Follow-up Interview are denoted as V10VISITYP=4 (Close-out follow-up interview). There are 53 participants who had the first page (page i) of the 96-month Close-out Follow-up Interview entered into the data system, and refused the interview. These participants have data for V10VISITYP, V10VISDYS, and V10AGE. All other variables are set to missing.

As of release 10.2.2, the 96-month self-administered weight loss variables from Dr. Thomas Link's approved OAI Ancillary Study #AS12-22 and the 96-month self-administered physical activity variables from Dr. Grace Lo's approved OAI Ancillary Study #AS11-20 have been released.

Although the category/subcategory organization of the data is independent of the grouping of variables into SAS datasets, the following category groupings are largely or entirely contained within these datasets:

- Bookkeeping/ all subcategories
- Demographics/ various subcategories
- Global function/disability/QOL/ Employment and work
- Health care access/ all subcategories
- Medical history, arthritis/ Family history
- Medical history, general/Height and weight history (96-month weight loss variables only)
- Physical activity/ all subcategories
- Study eligibility/ various subcategories

Data Structure/Contents:
These datasets contain one record per participant. The variable uniquely identifying a record is ID, and the datasets are sorted by ID, which can be used to merge/join to data in other datasets.

A full list and description of all the variables contained in these datasets can be found in the contents.pdf.

Condition of data:

- **Known data errors:** problems/cautions for use are listed by variable in the "Release Comments."
- **Dataset strengths/weaknesses:** None at this time.

General strategies for use:
When using with other datasets, merge/join by ID.

Analysts are encouraged to always output and view SAS variable labels in their entirety to ensure important information about the variables is not lost. The maximum SAS label length is 160 characters.