Get your suit, we’re going to the pool! Approaches for pooling across different datasets.

Briana Mezuk, PhD
MCUAAAR AnC Co-Director
February 19, 2020
bmezuk@umich.edu

The problem(s):

- Your scientific question relates to a process that occurs over the life course, but you only have data on people of a particular age
- The number of observations in your dataset is limited (sometimes recruitment doesn’t go as planned…)
- You have multiple datasets that have information relevant to your question, and you have a feeling there’s a way you can make that work to your advantage, but you don’t know how

So what does it mean to “pool” datasets?

- It means to take information you have on some people (i.e., CVD risk factors for younger people in one cohort) and apply it to a different set of people to “fill in” their values that you didn’t collect data on (i.e., CVD risk factors for people in another cohort who were older when data collection started for their cohort).

Q: What other situations do you know of when we take observed data and use it to “fill in” unobserved (missing) values?

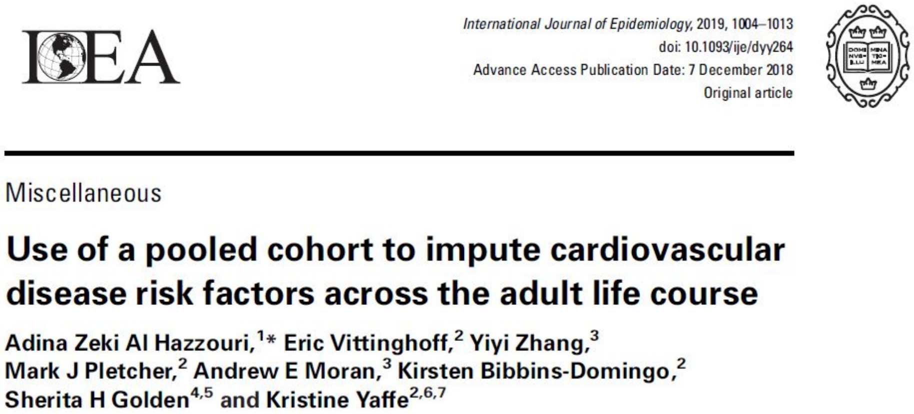
A: Multiple imputation!

Inspiration for today’s session

Problem: We think that risk for cardiovascular disease (CVD) starts early in life, but few single cohorts have followed participants over the entire etiologically-relevant period.

So what? This means that we have limited information on effective prevention strategies.

Solution: POOL COHORTS that represent a range of participants over the life span.



IJEA

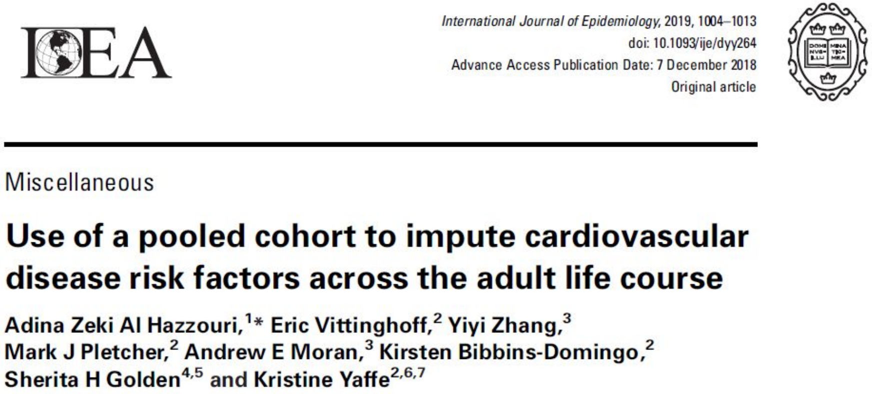
International Journal of Epidemiology, 2019, 1004–1013
doi: 10.1093/ije/dyy264
Advance Access Publication Date: 7 December 2018
Original article

Miscellaneous

Use of a pooled cohort to impute cardiovascular disease risk factors across the adult life course

Adina Zeki Al Hazzouri,[1]* Eric Vittinghoff,[2] Yiyi Zhang,[3] Mark J Pletcher,[2] Andrew E Moran,[3] Kirsten Bibbins-Domingo,[2] Sherita H Golden[4,5] and Kristine Yaffe[2,6,7]

Goal of this study



International Journal of Epidemiology, 2019, 1004–1013
doi: 10.1093/ije/dyy264
Advance Access Publication Date: 7 December 2018
Original article

Miscellaneous

Use of a pooled cohort to impute cardiovascular disease risk factors across the adult life course

Adina Zeki Al Hazzouri,[1]* Eric Vittinghoff,[2] Yiyi Zhang,[3] Mark J Pletcher,[2] Andrew E Moran,[3] Kirsten Bibbins-Domingo,[2] Sherita H Golden[4,5] and Kristine Yaffe[2,6,7]

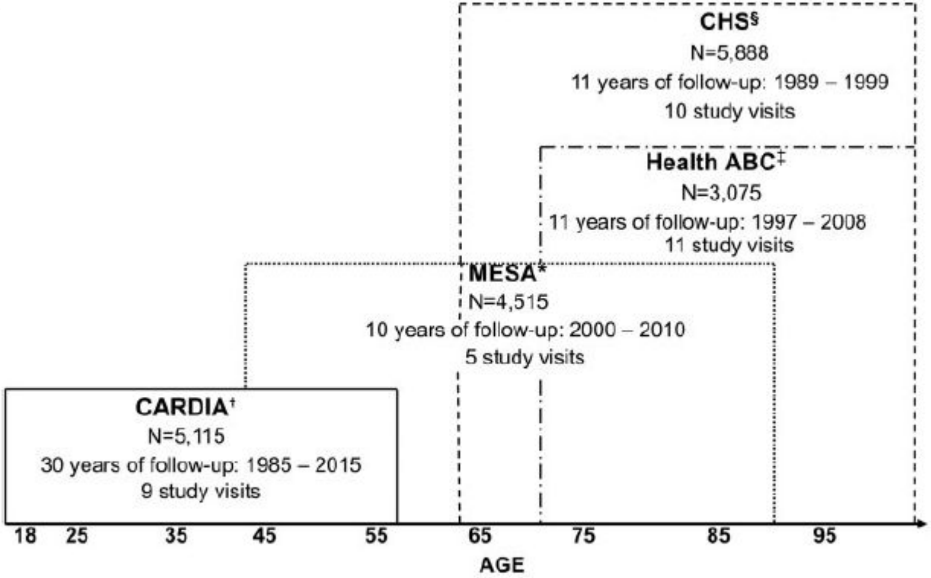
Aim: Use information from other cohorts, and what we know about the etiology of CVD, to impute early and midlife levels risk factors for CVD.

Lifecourse trajectories of CVD-related outcomes they wanted to impute: BMI, glucose, lipids, blood pressure.

Risk factors used to impute these outcome trajectories: smoking status, diabetes status, hypertension status, medication use for diabetes, hypertension, and high cholesterol.

Four cohorts covering different periods of the life course

1. Coronary Artery Risk Development in Young Adults (CARDIA)
2. Multi Ethnic Study of Atherosclerosis (MESA)
3. Cardiovascular Health Study (CHS)
4. Health, Aging and Body Composition (Health ABC) Study


CHS§
N=5,888
11 years of follow-up: 1989 – 1999
10 study visits
Health ABC‡
N=3,075
11 years of follow-up: 1997 – 2008
11 study visits
MESA*
N=4,515
10 years of follow-up: 2000 – 2010
5 study visits
CARDIA†
N=5,115
30 years of follow-up: 1985 – 2015
9 study visits
18
25
35
45
55
65
75
85
95
AGE

† Study visits took place at years 0, 2, 5, 7, 10, 15, 20, 25, and 30

* Study visits took place at years 0, 2, 4, 5, and 10

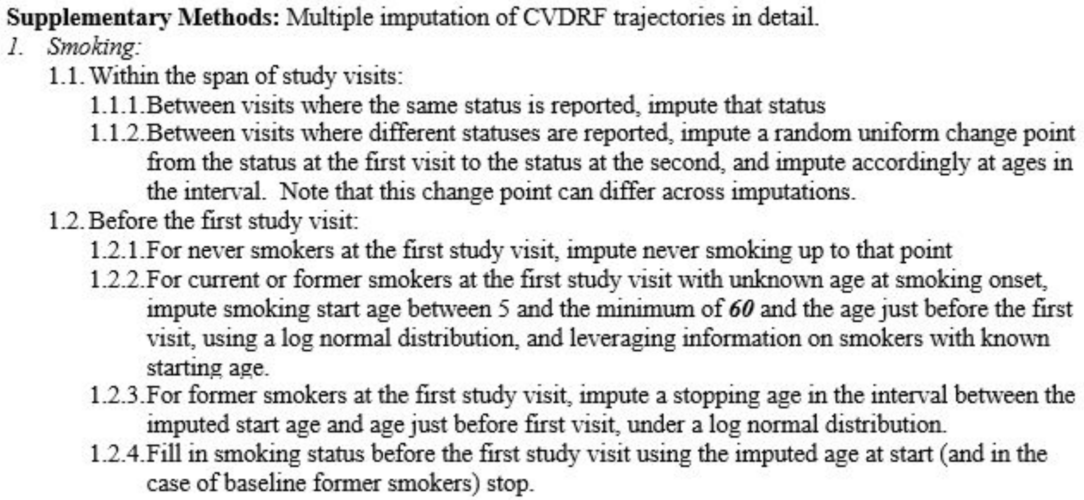
§ Study visits took place at years 2 (baseline), 3, 4, 5, 6, 7, 8, 9, 10, and 11

‡ Annual study visits

Figure 1. Description of the four study cohorts by age, sex and race.

Necessary ingredients for pooling

1. Similar, if not identical, measures across the cohorts, which you operationalize in an identical way across all cohorts
2. Logic for imputation that you can implement consistently across all cohorts
    a. Example here for how they imputed smoking status values across all cohorts.

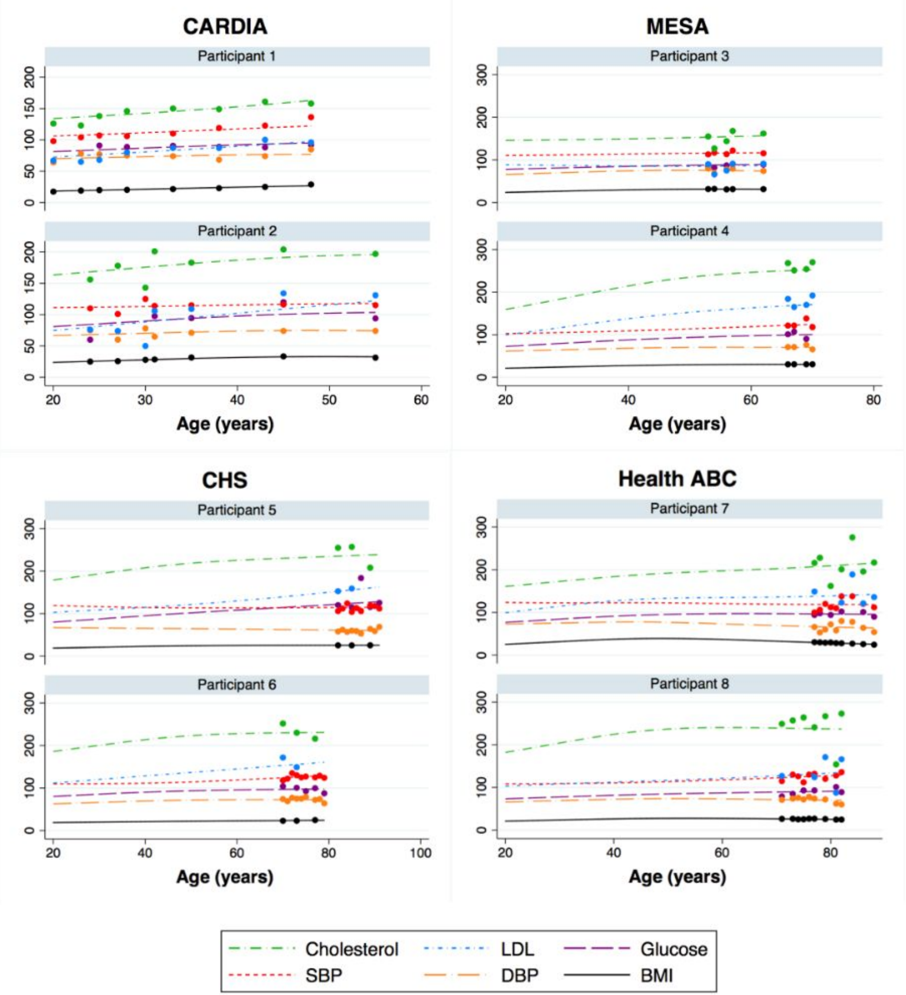


Supplementary Methods: Multiple imputation of CVDRF trajectories in detail.

1. *Smoking:*
    1.1. Within the span of study visits:
        1.1.1. Between visits where the same status is reported, impute that status
        1.1.2. Between visits where different statuses are reported, impute a random uniform change point from the status at the first visit to the status at the second, and impute accordingly at ages in the interval. Note that this change point can differ across imputations.
    1.2. Before the first study visit:
        1.2.1. For never smokers at the first study visit, impute never smoking up to that point
        1.2.2. For current or former smokers at the first study visit with unknown age at smoking onset, impute smoking start age between 5 and the minimum of ***60*** and the age just before the first visit, using a log normal distribution, and leveraging information on smokers with known starting age.
        1.2.3. For former smokers at the first study visit, impute a stopping age in the interval between the imputed start age and age just before first visit, under a log normal distribution.
        1.2.4. Fill in smoking status before the first study visit using the imputed age at start (and in the case of baseline former smokers) stop.

Observed data (dots) and imputed data (lines)

- **The main aim of this study** was to impute CVD risk indicators for earlier in the adult life course, beginning from age 20 years, *for participants with follow-up data beginning in mid-life or later in life.*
  - For example, for a MESA participant first seen at age 45 and last seen at age 55, the expected CVD risks were imputed each year from age 20 to 55.
- **Implication**: The amount of missing data imputed varied across the cohorts as a function of the age of participants at baseline.



How they pooled data from these cohorts to estimate life course trajectories of CVD risk

- CVD outcome trajectories were estimated using linear mixed models (LMMs).
    - **What is a mixed model? One that has both fixed and random effects.**
- And what are fixed and random effects…?
    - **A fixed effect is a parameter that does not vary**. That is, we assume there is some true regression line in the population (β) and we use our data to get some estimate of it (β-hat), which varies from sample to sample just because that's the way sampling probability works.
    - **A random effect is a parameter that is itself a random variable**. That is, we assume that the true mean (population) β is not a single "true" value, but instead assume that β itself has a normal distribution with mean (μ) and a standard deviation (σ).
- Nice resource on LMMs
  https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

Random effects used in this study:

- Intercept
- Age

How they pooled data from these cohorts to estimate life course trajectories of CVD risk

- These LMM were used to generate ***best linear unbiased predictions (BLUPs) for each person in the pooled data****,* annually from age 20-years until the end of follow-up for each participant.

- Using those BLUP trajectories, they calculated period-specific t**ime-weighted averages (TWAs) to summarize** early (ages 20–39 years) and midlife (ages 40–59 years) CVD risk factors using data from all cohorts.
  - “Time-weighted” means that the values are set at zero at ages before the interval, time-dependent within the interval, and fixed at the value from the last age in the interval for ages post-interval.

Overall time-weighted averages for the CVD risk indicators

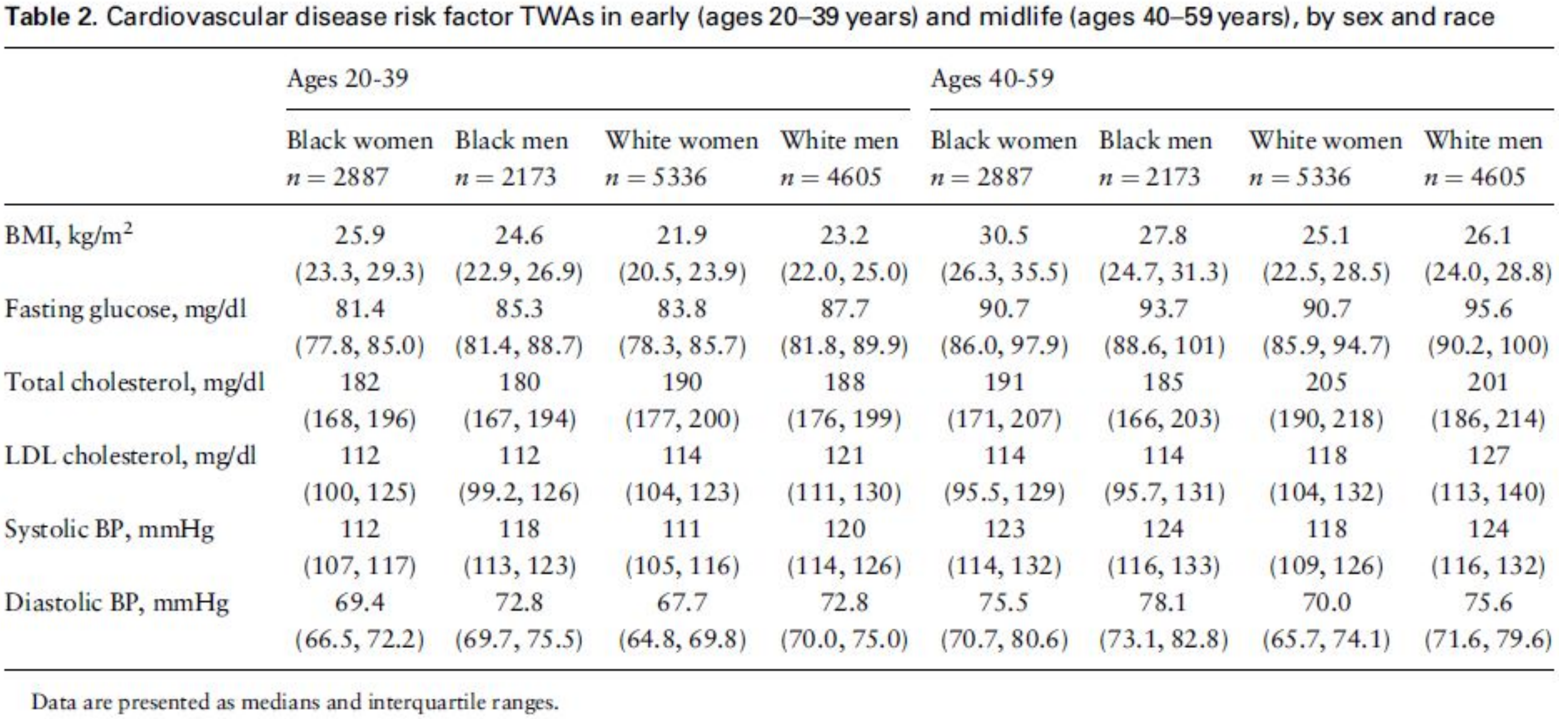
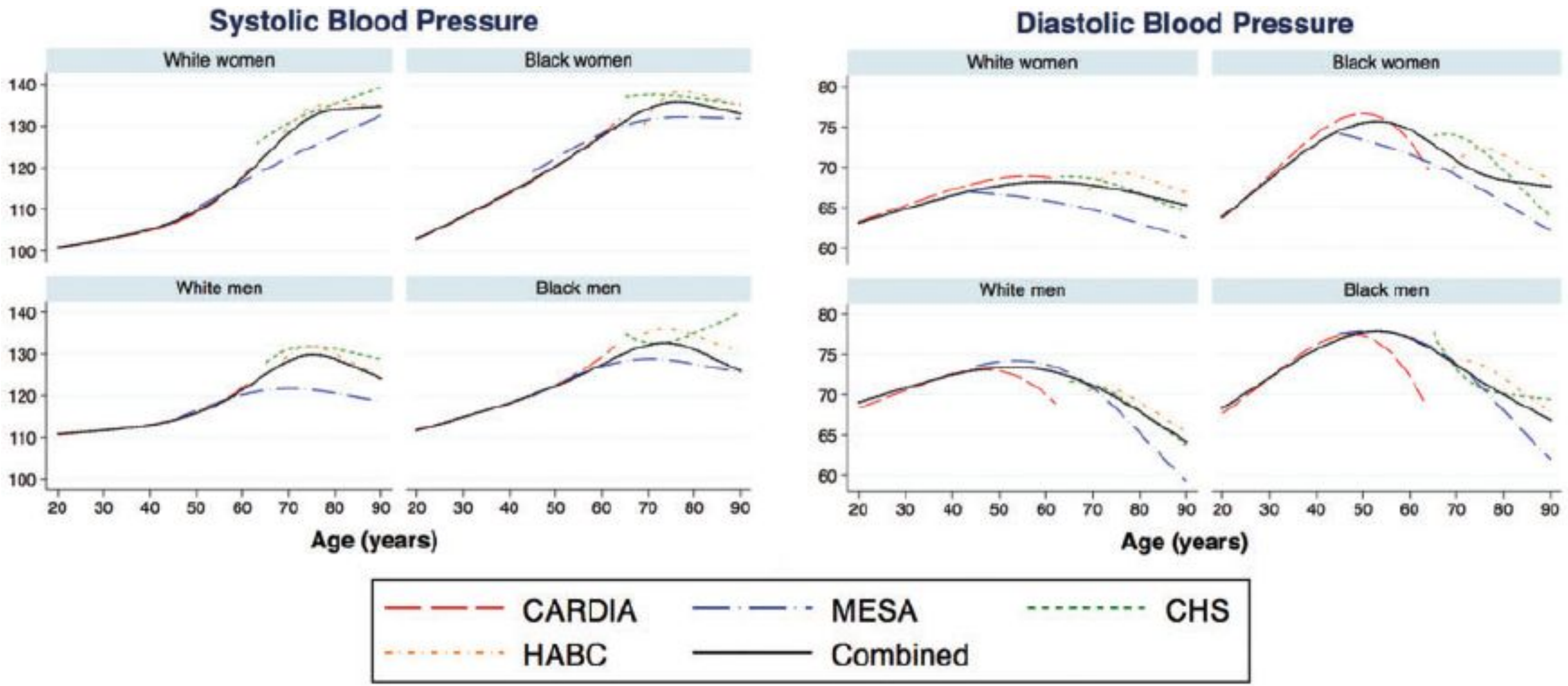


Table 2. Cardiovascular disease risk factor TWAs in early (ages 20–39 years) and midlife (ages 40–59 years), by sex and race

	Ages 20-39				Ages 40-59			
	Black women *n* = 2887	Black men *n* = 2173	White women *n* = 5336	White men *n* = 4605	Black women *n* = 2887	Black men *n* = 2173	White women *n* = 5336	White men *n* = 4605
BMI, kg/m^2	25.9 (23.3, 29.3)	24.6 (22.9, 26.9)	21.9 (20.5, 23.9)	23.2 (22.0, 25.0)	30.5 (26.3, 35.5)	27.8 (24.7, 31.3)	25.1 (22.5, 28.5)	26.1 (24.0, 28.8)
Fasting glucose, mg/dl	81.4 (77.8, 85.0)	85.3 (81.4, 88.7)	83.8 (78.3, 85.7)	87.7 (81.8, 89.9)	90.7 (86.0, 97.9)	93.7 (88.6, 101)	90.7 (85.9, 94.7)	95.6 (90.2, 100)
Total cholesterol, mg/dl	182 (168, 196)	180 (167, 194)	190 (177, 200)	188 (176, 199)	191 (171, 207)	185 (166, 203)	205 (190, 218)	201 (186, 214)
LDL cholesterol, mg/dl	112 (100, 125)	112 (99.2, 126)	114 (104, 123)	121 (111, 130)	114 (95.5, 129)	114 (95.7, 131)	118 (104, 132)	127 (113, 140)
Systolic BP, mmHg	112 (107, 117)	118 (113, 123)	111 (105, 116)	120 (114, 126)	123 (114, 132)	124 (116, 133)	118 (109, 126)	124 (116, 132)
Diastolic BP, mmHg	69.4 (66.5, 72.2)	72.8 (69.7, 75.5)	67.7 (64.8, 69.8)	72.8 (70.0, 75.0)	75.5 (70.7, 80.6)	78.1 (73.1, 82.8)	70.0 (65.7, 74.1)	75.6 (71.6, 79.6)

Data are presented as medians and interquartile ranges.

Life course trajectories of cardiovascular disease risk factors across categories of sex and race, results from the specific cohorts and pooled cohort.



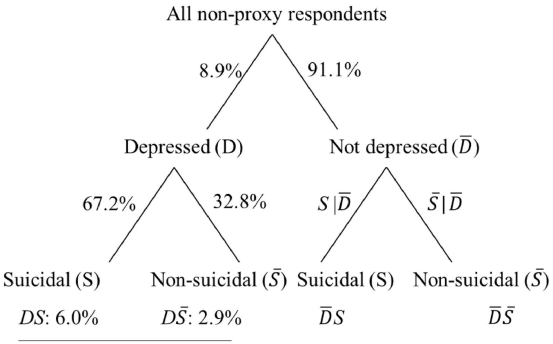
Other types of pools: Prediction modeling using an external dataset

- **Step 1:** Find another dataset that has data on similar measures as the ones that you are hoping to “fill in” information on.
- **Step 2**: Estimate a logistic regression model in the external dataset to generate beta-coefficients for a “prediction model” that you will use to estimate the missing values in your data.
- **Step 3**: Append the external dataset to your dataset.
- **Step 4**: Fit the prediction model on the entire sample (actual + external) and use the predicted probabilities from the logistic regression model to “fill in” the missing data in your primary data.

Other types of pools:
Prediction modeling using an external dataset

- **What type of problem does this solve?**
  - You have systematic missing data on a variable of interest and don’t feel that other missing data approaches (e.g., multiple imputation) are appropriate.
  - Ex. Some of your respondents were simply not asked a question because of a skip pattern in the interview.
- **When is it an appropriate solution?**
  - When you have an external dataset that is sufficiently similar to the one you are working with in terms of population, scope, and - most importantly - measures, but doesn’t have the same missing data problem as your dataset.

Example:
Prediction modeling using an external dataset



All non-proxy respondents

8.9%

91.1%

Depressed (D)

Not depressed ($\bar{D}$)

67.2%

32.8%

$S \mid \bar{D}$

$\bar{S} \mid \bar{D}$

Suicidal (S)

Non-suicidal ($\bar{S}$)

Suicidal (S)

Non-suicidal ($\bar{S}$)

DS: 6.0%

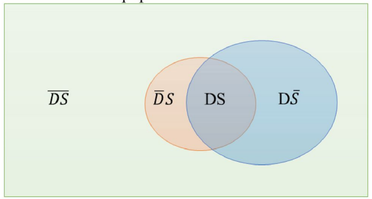
$D\bar{S}$: 2.9%

$\bar{D}S$

$\bar{D}\bar{S}$

Overall prevalence of ideation in the HRS based on observed data only

Heuristic Venn diagram of the conceptual distinction between depression and suicidal ideation in the population overall

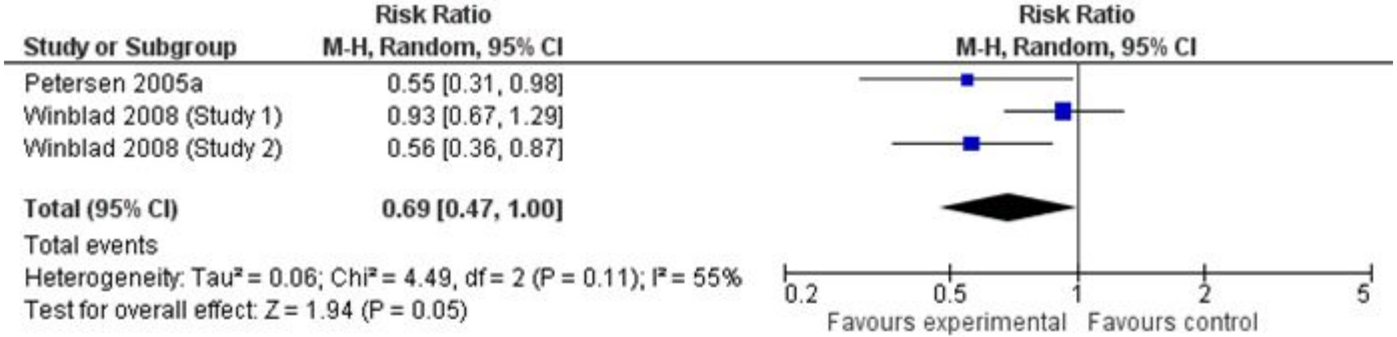


- **The Question**: What is the prevalence of passive suicidal ideation among older adults in the US?
- **The Problem:** The HRS only asks participants the item on suicidal ideation if they first endorse feelings of depression (skip pattern).
- **Solution**:
  1. Find an external dataset that doesn't have this skip pattern: Baltimore ECA
  2. Use it to empirically estimate the predictors of suicidal ideation among people who are not depressed
  3. Append the ECA to the HRS and apply those beta-coefficients back to the HRS to estimate the size of the **-D/+S** group.

Other types of pools

Cochrane

Meta-analysis: is a quantitative, formal, epidemiological study design used to systematically assess the results of previous research to derive conclusions about that body of research.



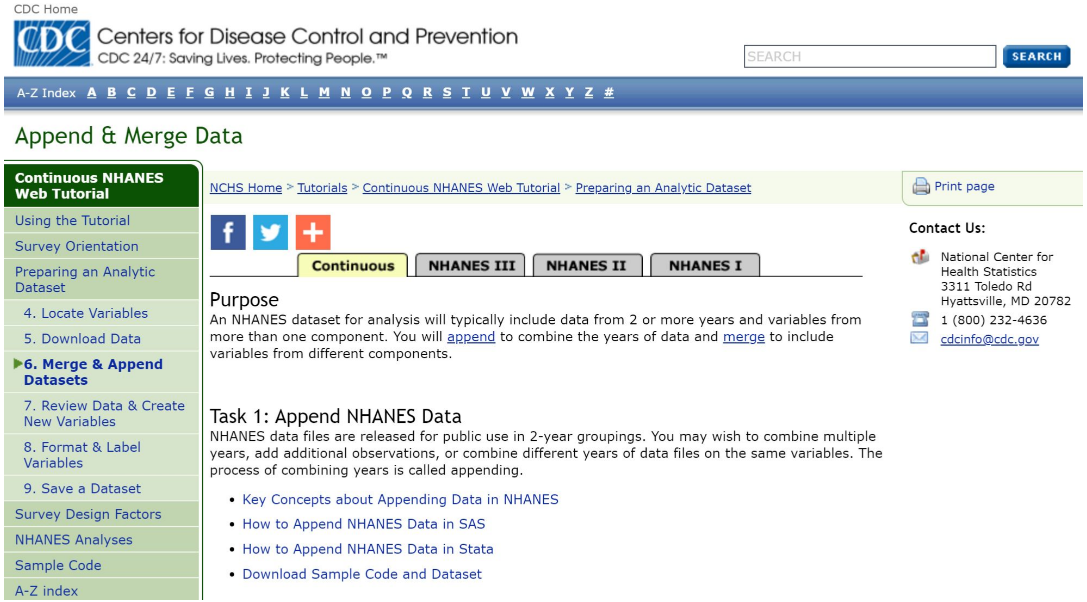
Study or Subgroup	Risk Ratio M-H, Random, 95% CI
Petersen 2005a	0.55 [0.31, 0.98]
Winblad 2008 (Study 1)	0.93 [0.67, 1.29]
Winblad 2008 (Study 2)	0.56 [0.36, 0.87]
Total (95% CI)	**0.69 [0.47, 1.00]**
Total events	

Heterogeneity: $Tau^2 = 0.06$; $Chi^2 = 4.49$, df = 2 ($P = 0.11$); $I^2 = 55\%$

Test for overall effect: $Z = 1.94$ ($P = 0.05$)

Appending cross-sectional panels as a way to pool

- Some data are designed for pooling! NHANES has specific instructions for how to do this.
- Why? To increase statistical power, particularly if your question concerns smaller demographic subgroups


CDC Home
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™
SEARCH
SEARCH
A-Z Index A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #
Continuous NHANES Web Tutorial
Using the Tutorial
Survey Orientation
Preparing an Analytic Dataset
4. Locate Variables
5. Download Data
6. Merge & Append Datasets
7. Review Data & Create New Variables
8. Format & Label Variables
9. Save a Dataset
Survey Design Factors
NHANES Analyses
Sample Code
A-Z index
NCHS Home > Tutorials > Continuous NHANES Web Tutorial > Preparing an Analytic Dataset
Print page
Continuous
NHANES III
NHANES II
NHANES I

Append & Merge Data

Purpose

An NHANES dataset for analysis will typically include data from 2 or more years and variables from more than one component. You will append to combine the years of data and merge to include variables from different components.

Task 1: Append NHANES Data

NHANES data files are released for public use in 2-year groupings. You may wish to combine multiple years, add additional observations, or combine different years of data files on the same variables. The process of combining years is called appending.

- Key Concepts about Appending Data in NHANES
- How to Append NHANES Data in SAS
- How to Append NHANES Data in Stata
- Download Sample Code and Dataset

Contact Us:

National Center for Health Statistics
3311 Toledo Rd
Hyattsville, MD 20782

1 (800) 232-4636

cdcinfo@cdc.gov

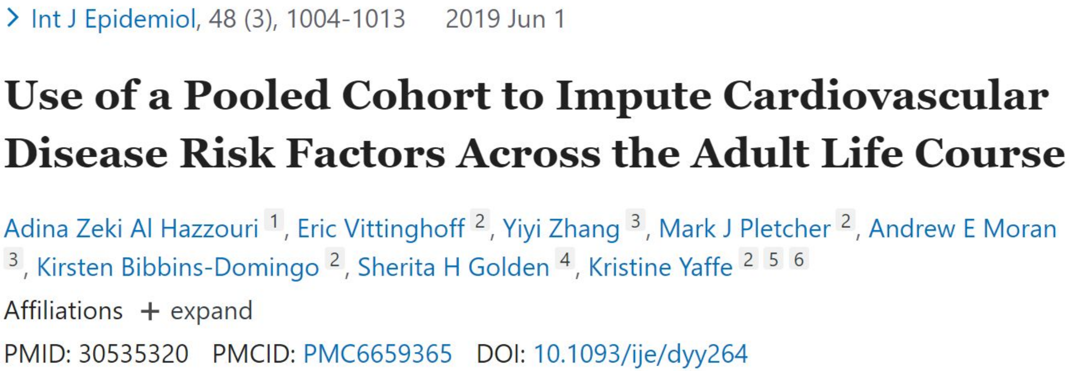
Types of problems pooling does not solve

- Differential attrition (healthier people are more likely to stay in your study over time)
- Survival bias (healthier people more likely to make it to older age, where your study begins)
- May inadvertently “smooth out” meaningful variation

Summary

- The fundamental idea underlying pooling datasets is that you can take information in one study and use it to “fill in the gaps” in another, related study.
- There are several variations on this theme, but they all stem from the idea that **the data you have can be used to impute the data you do not.**
- Like all imputation strategies, these approaches can address some sources of bias but at a cost of precision.

Further reading...



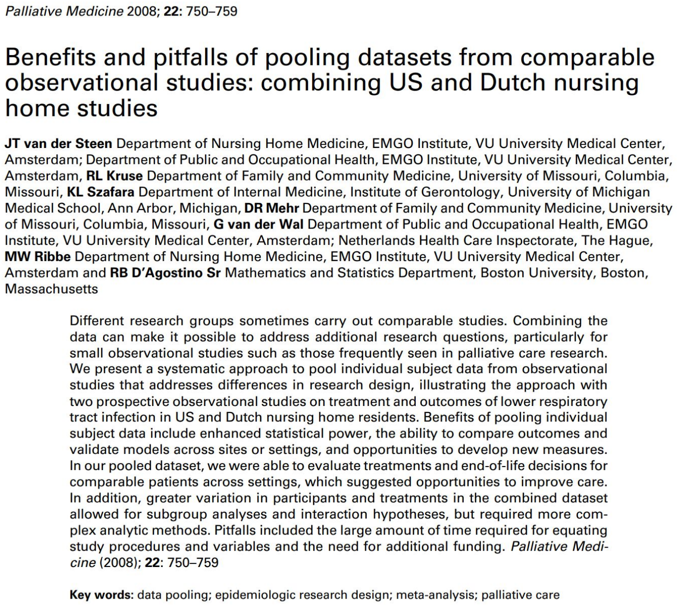
> Int J Epidemiol, 48 (3), 1004-1013   2019 Jun 1

Use of a Pooled Cohort to Impute Cardiovascular Disease Risk Factors Across the Adult Life Course

Adina Zeki Al Hazzouri [1], Eric Vittinghoff [2], Yiyi Zhang [3], Mark J Pletcher [2], Andrew E Moran [3], Kirsten Bibbins-Domingo [2], Sherita H Golden [4], Kristine Yaffe [2] [5] [6]

Affiliations + expand

PMID: 30535320   PMCID: PMC6659365   DOI: 10.1093/ije/dyy264



Palliative Medicine 2008; **22:** 750–759

Benefits and pitfalls of pooling datasets from comparable observational studies: combining US and Dutch nursing home studies

JT van der Steen Department of Nursing Home Medicine, EMGO Institute, VU University Medical Center, Amsterdam; Department of Public and Occupational Health, EMGO Institute, VU University Medical Center, Amsterdam, **RL Kruse** Department of Family and Community Medicine, University of Missouri, Columbia, Missouri, **KL Szafara** Department of Internal Medicine, Institute of Gerontology, University of Michigan Medical School, Ann Arbor, Michigan, **DR Mehr** Department of Family and Community Medicine, University of Missouri, Columbia, Missouri, **G van der Wal** Department of Public and Occupational Health, EMGO Institute, VU University Medical Center, Amsterdam; Netherlands Health Care Inspectorate, The Hague, **MW Ribbe** Department of Nursing Home Medicine, EMGO Institute, VU University Medical Center, Amsterdam and **RB D'Agostino Sr** Mathematics and Statistics Department, Boston University, Boston, Massachusetts

Different research groups sometimes carry out comparable studies. Combining the data can make it possible to address additional research questions, particularly for small observational studies such as those frequently seen in palliative care research. We present a systematic approach to pool individual subject data from observational studies that addresses differences in research design, illustrating the approach with two prospective observational studies on treatment and outcomes of lower respiratory tract infection in US and Dutch nursing home residents. Benefits of pooling individual subject data include enhanced statistical power, the ability to compare outcomes and validate models across sites or settings, and opportunities to develop new measures. In our pooled dataset, we were able to evaluate treatments and end-of-life decisions for comparable patients across settings, which suggested opportunities to improve care. In addition, greater variation in participants and treatments in the combined dataset allowed for subgroup analyses and interaction hypotheses, but required more complex analytic methods. Pitfalls included the large amount of time required for equating study procedures and variables and the need for additional funding. *Palliative Medicine* (2008); **22:** 750–759

Key words: data pooling; epidemiologic research design; meta-analysis; palliative care