VITAMIN D

TABLE OF CONTENTS

Introduction 2

Date of Blood Draw Collection 2

Matching 2

Sampling Techniques 2

Breast and Prostate 2

Pancreas 2

Endometrial, Ovarian, Renal and Upper GI 3

Bladder 3

Frozen Data 3

Other Data Availability 3

VITAMIN D

INTRODUCTION

The primary vitamin D dataset combines the results of thirteen separate studies. The chronologically first measurement was taken when participants were in multiple studies. To determine the source of the measurement, use the variable VITD_STUDY. Within each individual study dataset are the identifiers, case/control indicators, the cell or case set groupings, the matching variables used at the time, specimen chronology, and the vitamin D results.

DATE OF BLOOD DRAW COLLECTION

In PLCO dates are considered sensitive personal information. In all PLCO datasets on CDAS, all date variables are presented as days variables normalized around randomization, with the year of randomization given as the point of reference. Since vitamin D has seasonal variation, most of the studies matched on date of blood draw in some form. Though PLCO is unable to provide these dates, the season of blood draw is provided along with the calendar year. For the purposes of these datasets, the seasons run from December through February, March through May, June through August and September through November. Any given calendar year begins on December 1st instead of January 1st.

MATCHING

SAMPLING TECHNIQUES

Some of the studies were selected with frequency matching and some used individual matching. When cases were frequency matched to controls, the variable CELL is included which pairs all cases within a set of matching strata with their controls. If a study used individual matching, the variable CASE_SET is included, which is unique per case and its matched control. All matching variables are included within each dataset, except any which pertain to date of blood draw due to issues mentioned above. Some studies utilized replacement, meaning that a control could be selected more than once, and cases could also be selected as controls prior to their diagnosis. Within any given study, only the latest instance of selection is included.

BREAST AND PROSTATE

Both the breast and prostate studies were selected as subsets of already matched cohorts from the Cancer Genetic Markers of Susceptibility (CGEMS) study. The matching variables and cells were retained from the original study; however the case and control counts within each cell will not match.

PANCREAS

The pancreatic cancer study was handled as two separate selections. For both selections, neo-endocrine cases were excluded. The first selection used month and year of randomization for matching. Later it became clear that, while they happened very close in proximity, this event wasn’t always done simultaneously with the T0 blood draw. Thus, for the second phase of the study, the month and year of the blood draw were used instead. CELL retains the matching factor that was used at the time.

Endometrial, Ovarian, Renal and Upper GI

These populations were all selected simultaneously. Any participant that overlapped between these studies was only measured once. All matching factors are identical, with the only difference in selection being gender and surgery considerations for the exclusions in the endometrial and ovarian populations.

Bladder

When the bladder population was selected, matching was done on study year of blood draw, and the date of that blood draw. For 23 control subjects who had been selected to match cases with a T1 or T2 blood draw, the T0 blood was used instead. Both the study year on which they were matched as well as the study year that was measured are provided. The specimen chronology variables all reflect the study year that was measured.

Frozen Data

The populations for these studies were selected at different time points ranging from May 2001 to May 2015. They are thus frozen in time. Since the time of selection and the time of analysis, some of the controls will have been diagnosed with cancer. The variable IS_CASE reflects the case status at the time of selection, though current cancer status can be ascertained from the various site datasets on CDAS. For the breast and pancreas studies, in order to maximize the number of cases, participants who had reported a cancer to the trial, but whose medical records had not yet confirmed it, were included in the study. If a cancer was deemed erroneously reported, or the records could not be obtained, this case was removed from the population, and in the pancreas study, their matched controls were removed as well. The final denominators in these files represent everybody who was selected for the study, who still meet eligibility for the study, and who have a vitamin D result available.

Other Data Availability

As it was alluded to above, several of these populations were selected in concert with other EEMS populations, and thus these populations may be good candidates for other research. The breast, prostate, pancreas and bladder sets all overlap heavily with the PLCO genome-wide association studies (GWAS) populations. Various other biochemical and chromosomal analytes have also been measured in these same groups of cases and controls, though they are not currently available on CDAS.