Supplemental File: Approximation of medication use history in WHI cohorts

Medication inventories were collected during the WHI study at baseline and years 1, 3, 6, and 9 in the CT, and at baseline and year 3 in the OS (Figure 1). At each inventory, clinic interviewers entered each medication name and strength directly from the containers into a database that assigned drug codes using Medi-Span software (First DataBank, Inc., San Bruno, CA). The duration of each medication, based on the participant's recollection, was also entered. Similar medication use data was also collected, by mail or phone, towards the end of the first WHI extension period. The collective data was then used to construct an approximation to each participant's medication use history over time, and subsequently used as a time-dependent exposure variable in our manuscript. This approximation and rationale is described below.

A basic approximation of medication use consists of representing a participant's history by multiple rows of data; each row represents the non-overlapping follow-up after a collection. For example, the medications use history of a CT participant during the WHI study for a particular drug could be approximated by five rows of data: (0, Y1,Z(0)], (Y1,Y3,Z(1)], (Y3,Y6,Z(3)], (Y6,Y9,Z(6)] and (Y9,T,Z(9)]. Here Yi = time from randomization to the ith collection, Z(i) is an indicator variable(s) for medication use, and T is the time to event or censoring. Data in this *counting process* format can readily be analyzed in a *Cox proportional hazards* regression model that allow for time-dependent variables such as Z(i).

An obvious consequence of the basic approximation is that Z(i) is not updated until the next collection, at annual visit i+1. This limitation may not lead to a reliable approximation for medications whose usage pattern changed markedly over time. For example, 32% of the participants being treated for diabetes were using Metformin at year 1, and the proportion of those treated with metformin increased to 45% by year 3 (Figure 2). While it is very likely that participants not using metformin at year 1, later initiated metformin during the interim, the basic approximation does not address changes in drug use *between* visits. Rather the basic approximation

assumes initiation of metformin at year 3. Due to this major limitation, the basic approximation was not used in our paper.

A better approximation can be obtained by using duration data to refine the time of initiation during follow-up. Specifically, the refined time of initiation is computed as Yi minus the minimum (δ, duration reported at Yi), where $\delta = 1$ for year 3, and $\delta = 2$ for years 6 and 9. Hence the refined initiation times would lie in the intervals (*Y2*, Y3], (*Y4*, Y6], and (*Y7*, Y9], respectively. Limiting the refined initiation time to the aforementioned intervals, extends a participant's recall of medication use to at most one (two) year(s), and consequently does not supersede the preceding collection. A similar refinement can be used at year 3 for the OS where $\delta = 2$. To ensure a similarly reliable approximation of a participant's medication use history during the extension, time to initiation should be refined by no more than two years.

Another refinement can be made by accounting for collection data that has become out-of-date, and thereby no longer accurately represents a participant's medication use. Due to the collection design (Figure 1), we chose to censor the follow-up for a participant 3.5 years after the last collection; 3 years per design and allowing 0.5 years for sample variability. For example, if a CT participant missed her year 6 collection, the history could be approximated by *discontinuous intervals of risk:* (0, Y1,Z(0)], (Y1,Y3,Z(1)], (Y3, Y3 + 3.5, Z(3)], and (Y9,T,Z(9)]. The participant's history during the interval (Y3 + 3.5, Y9] would not be included in the analysis, although the participant would be allowed to re-enter the analysis at year 9.

A brief comparison of the approximations to participants' medication use histories over time is shown in Figure 2. The proportion of metformin users among diabetics treated with known medications is shown at each follow-up visit including the extension. The usage pattern of metformin increases markedly during the study period. It is very likely, that the true proportion of metformin users is grossly underestimated by the basic approximation at year 2; many OS

participants may have initiated metformin use before their year 3 collection. Assuming the proportion of the *full* cohort using metformin grew linearly between years 1 and 3 (pink line), the basic approximation is greatly improved, between years 2 and 3, by leveraging duration data to refine the time of initiation. The approximation that censors out-of-date medications improves upon the basic approximation after 3.5 years from enrollment, with further improvements 6.5 and 9.5 years after enrollment. Censoring out-of-date medications resulted in marked improvement to the basic approximation; integrated squared error (ISE) was reduced by half[1]. The approximation used in our manuscript, that also refines initiation time, reduced ISE by 85%.

A final refinement was made to prevent differential follow-up, during the extension period, between participants that were or were not using anti-diabetic medication. While medication use histories were collected towards the end of the first five year *extension*, for most WHI CT and OS participants, the most recent *study* collection occurred at Y6 and Y3, respectively. Consequently, medication histories collected during the study became out-of-date and were censored, until re-entry into the analysis with more recent extension data. To prevent differential follow-up, participants not taking the medications of interest were allowed to re-enter the risk set two years prior to the date of their extension collection. For participants taking the medications of interest, initiation was refined by the minimum of 2 years, or the reported duration. If self-reported duration was less than 2 years, the remainder of the 2 year interval did not reflect using the medication(s) of interest.

[1] Assumes the subsample of participants with a medications collection is representative of the full cohort, and the usage pattern of the full cohort changes linearly between collections.

Figure Legends.

Figure 1: Medication collection design and refinements.

The top and bottom panels show the medication collections for the OS and CT, respectively. The vertical bars indicate when medications were collected during the study (blue) and the extension period (black). The heights of the vertical bars indicate what fraction of women, relative to baseline, had a collection. For example, a much smaller % of women in the CT had a collection at year 9. The shaded regions indicate variability when each annual visit occurred (+/- 1 SD). Left-pointing green arrows indicate the extent to which duration information, for a particular collection year, could be used to refine time to medication initiation. For example, refinement would be the minimum (δ, reported duration), where δ = one for year 3 and two years for years 6 and 9. The right-pointing arrows indicate when a recent medication collection is considered reliable (green) and when it is considered out-of-date (red dotted line w/ x); for simplicity, similar arrows that indicate out-of-date medication use during the extension are not shown.

Figure 2: A comparison of approximations to participants' medication use history over time.

Red circles indicate the proportion of metformin users (among diabetics with known medications) based on the collected medication inventory. Blue (dotted), green and black lines are estimates of the proportion of metformin users that were computed using the basic approximation, approximation that censors out-of-date medications, and the approximation that also refines initiation, respectively. The pink line interpolates the observed data, collected on a subsample of the cohort, and is assumed to be the true proportion of metformin use of the full cohort. The integrated squared error is the squared difference between an approximation (blue, green or black lines) and the linear interpolated estimate between each collection (truth; pink line), integrated between 0 and 16 years.

Figure 1.

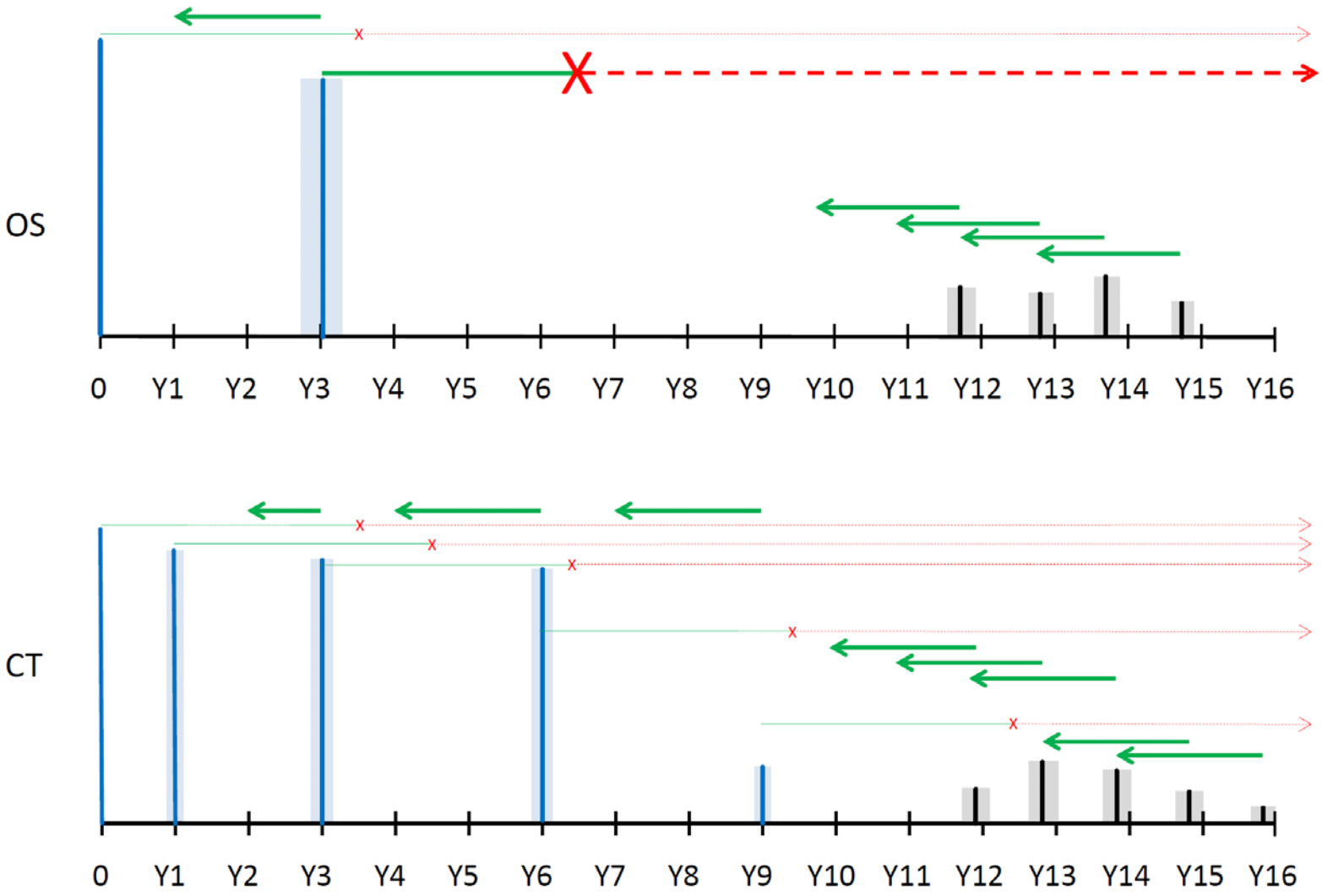
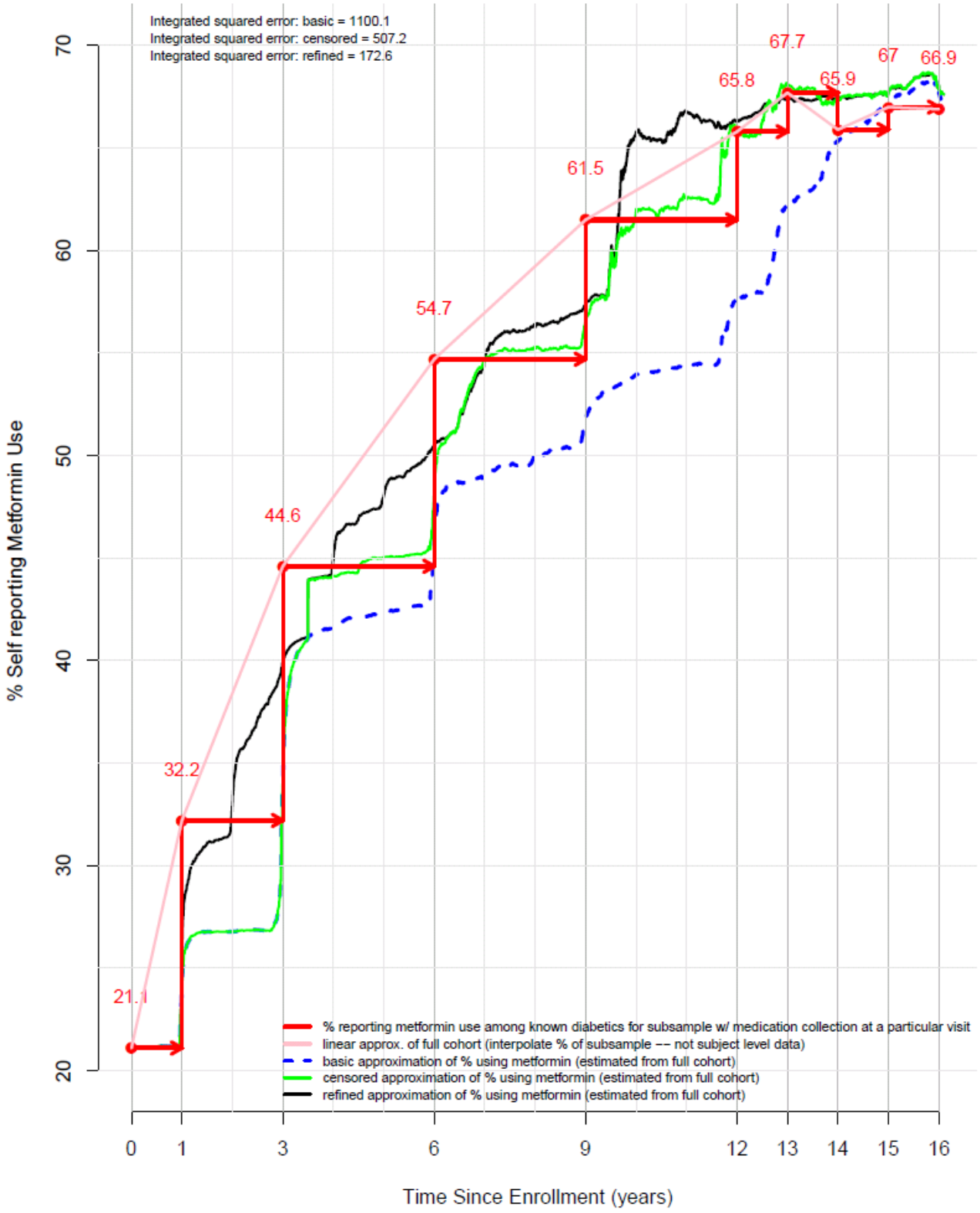

OS
0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Y10
Y11
Y12
Y13
Y14
Y15
Y16
CT
0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Y10
Y11
Y12
Y13
Y14
Y15
Y16

Figure 2.


Integrated squared error: basic = 1100.1
Integrated squared error: censored = 507.2
Integrated squared error: refined = 172.6
21.1
32.2
44.6
54.7
61.5
65.8
67.7
65.9
67
66.9
% Self reporting Metformin Use
20
30
40
50
60
70
0
1
3
6
9
12
13
14
15
16
Time Since Enrollment (years)
% reporting metformin use among known diabetics for subsample w/ medication collection at a particular visit
linear approx. of full cohort (interpolate % of subsample -- not subject level data)
basic approximation of % using metformin (estimated from full cohort)
censored approximation of % using metformin (estimated from full cohort)
refined approximation of % using metformin (estimated from full cohort)