

# Time-to-Event Estimation with Unreliably Reported Events in Medicare Health Plan Payment

Oana M. Enache<sup>a,\*</sup>, Sherri Rose<sup>b</sup>

<sup>a</sup>*Stanford University School of Medicine, Department of Biomedical Data Science, Edwards Building, 300 Pasteur Drive, Stanford, CA, 94304*

<sup>b</sup>*Stanford University, Department of Health Policy, Encina Commons, 615 Crothers Way, Stanford, CA, 94305*

---

## Abstract

Time-to-event estimation (i.e., survival analysis) is common in health research, most often using methods that assume proportional hazards and no competing risks. Because both assumptions are frequently invalid, estimators more aligned with real-world settings have been proposed. An effect can be estimated as the difference in areas below the cumulative incidence functions of two groups up to a pre-specified time point. This approach, restricted mean time lost (RMTL), can be used in settings with competing risks as well. We extend RMTL estimation for use in an understudied health policy application in Medicare. Medicare currently supports healthcare payment for over 69 million beneficiaries, most of whom are enrolled in Medicare Advantage plans and receive insurance from private insurers. These insurers are prospectively paid by the federal government for each of their beneficiaries' anticipated health needs using an ordinary least squares linear regression algorithm. As all coefficients are positive and predictor variables are largely insurer-submitted health conditions, insurers are incentivized to upcode, or report more diagnoses than may be accurate. Such gaming is projected to cost the federal government \$40 billion in 2025 alone without clear benefit to beneficiaries. We propose several novel estimators of coding intensity and possible upcoding in Medicare Advantage, including accounting for unreliable reporting. We demonstrate estimator performance in simulated data leveraging the National Institutes of Health's All of Us study and also develop an open source R package to simulate realistic labeled upcoding data, which were not previously available.

*Keywords:* survival analysis, restricted mean time lost, restricted mean survival time, Medicare Advantage, upcoding, risk adjustment

---

## 1. Introduction

Time-to-event (TTE) analyses typically aim to compare differences between groups drawn from two or more populations in a randomized clinical trial or observational study. This may

---

\*Corresponding author

Email address: oenache@stanford.edu (Oana M. Enache)

include plotting events over time using either Kaplan-Meier or cumulative incidence curves. Most commonly, hazard ratios are used to estimate the magnitude of the effect between groups<sup>1–3</sup>. For example, a review of 66 clinical trials with TTE primary outcomes across four major medical journals found that 80% of trials reported a hazard ratio as a main finding, while only 21% of studies reported any alternative approaches<sup>4</sup>.

However, hazard ratios rely on a proportional hazards assumption, which is often unrealistic in health data<sup>1,4,5</sup>. Hazard ratios also have an unintuitive interpretation for nonstatistical audiences and are a relative value with unclear meaning for decision-making<sup>2,6</sup>. Less frequently, a difference or ratio of survival probabilities at a single time point (e.g., mean or median) is reported and compared across groups. These types of measures also have limitations in practice, including that they omit most of the data and may not be estimable in certain scenarios<sup>2,7</sup>.

Furthermore, alternate approaches are needed if competing risks, or when there is more than one mutually exclusive outcome, are present. Despite being pervasive in health studies, competing risks are frequently ignored, which can result in biased estimates of the primary outcome<sup>8–10</sup>. Additional important considerations in TTE analyses include the choice of monitoring period (i.e., the time window analyzed between a pre-specified origin and end time) and comparison group.

### *1.1. Restricted mean survival time-based methods*

Although first proposed in 1949<sup>11</sup>, restricted mean survival time (RMST) was revisited for TTE estimation in more contemporary literature<sup>5,6,12–14</sup>. RMST is defined as the area under the survival curve of time to an event for a single monitoring period. It can be interpreted as the mean time to event for all study participants followed in that monitoring period<sup>6</sup>. RMST addresses many of the issues of more popular comparison approaches. With large enough sample sizes, it is estimable nonparametrically, does not require proportional hazards, and is censoring independent<sup>6</sup>. Recent methods development has focused on expansions for adaptive and group sequential trials<sup>15,16</sup>, estimating RMST with varying end times<sup>7,17,18</sup>, and permutation testing for small sample sizes<sup>19–22</sup>.

In TTE analyses with a single outcome, the area above the survival curve in one monitoring period (or, the end time minus the RMST) is the restricted mean time lost (RMTL)<sup>23</sup>. RMTL has a number of appealing features, including that it can straightforwardly be extended to correspond to the area below cause- or event-specific cumulative incidence curves in competing risk settings<sup>10,23,24</sup>. In such scenarios, an event-specific RMTL corresponds to the area below its cumulative incidence curve and can be interpreted as the mean time without that particular event in a monitoring period<sup>10</sup>, or a summary of both how many and when events occur in that time. Differences in RMTL between two groups (for example a treatment and control or other comparison group) can also be used to quantify effects<sup>25</sup>.

### *1.2. Payment formulas in Medicare Advantage*

Our motivating application focuses on the evaluation of incident diagnostic coding in Medicare. Medicare is a federally funded insurance program administered by the Centers for Medicare and Medicaid Services (CMS), supporting over 69 million Americans who are aged

65 years and older or chronically disabled<sup>26</sup>. Beneficiaries choose between receiving coverage through Traditional Medicare (TM) or Medicare Advantage (MA). TM beneficiaries' care is generally paid directly by CMS for each service provided (often referred to as fee-for-service). In contrast, more than half of all Medicare beneficiaries are part of the MA program and receive health insurance from private insurers who are paid by CMS for each beneficiary's anticipated health needs.

We examine the Medicare risk adjustment algorithm that determines prospective health plan payments in MA based on binary demographic and health diagnostic variables using ordinary least squares linear regression. The outcome of this algorithm is a risk score, which is used to adjust a benchmark payment amount per beneficiary. Importantly, insurers retain the amount of money paid by CMS regardless of what care their beneficiaries actually receive<sup>27,28</sup>.

There are 115 variables that report the purported presence or absence of certain health conditions, termed hierarchical condition categories (HCCs), in the current risk adjustment formula<sup>29</sup>. Although HCCs overall correspond to dozens of distinct health conditions, certain subsets of these variables also represent severity levels within the same condition and are therefore billed in a mutually exclusive manner. For example, an insurer can only be paid for coding one of Pancreas Transplant Status, Diabetes with Severe Acute Complications, Diabetes with Chronic Complications, or Diabetes with Glycemic, Unspecified, or No Complications<sup>28</sup>. We propose that such sets of HCCs can be considered conceptually analogous to competing risks in TTE analyses because only one HCC can be recorded and paid for at a time, where the amount paid to insurers increases with severity level.

Evaluation of coding in MA is of great interest to health policy researchers and policymakers. The structure of the risk adjustment formula (called the CMS-HCC formula) as an ordinary least squares regression with only positive coefficients incentivizes insurers to code for as many diagnoses as possible to increase profits<sup>30-32</sup>. Insurers making their beneficiaries appear sicker than they are by coding for more diagnoses or more severe diagnoses than necessary is referred to as upcoding. We focus on two types of upcoding in this paper. The first we call severity-based upcoding, as beneficiaries who have a lower-severity version of an HCC are instead coded with a more severe HCC of that health condition. The second we refer to as any-available upcoding, where any beneficiaries previously not coded with a given HCC could be upcoded, potentially fraudulently. When upcoding has not been concretely confirmed (as is most often the case in real-world settings) we reflect this by describing it as possible upcoding.

Assessment of upcoding in MA is usually done by comparing frequency of coding of beneficiaries, or coding intensity, to coding of similar health conditions in TM, with slight variation on inclusion and exclusion criteria used. However, TM is known to have underreporting (i.e., undercoding) of health conditions<sup>28,33</sup>. This type of unreliable reporting is important to account for as measures of differences between the programs will likely be inflated otherwise. Alternative reference data besides TM diagnoses have been explored in prior work but remain underutilized, including mortality data<sup>34</sup> and prescription drug utilization<sup>35</sup>. More recently, a set of measures were proposed to distinguish new (incident) versus continuing (persistent) coding for individual HCCs<sup>36</sup>.

In 2025, high MA coding intensity—which likely includes upcoding—is estimated to cost CMS \$40 billion in unnecessary spending to private insurers without clear benefit to MA beneficiaries<sup>32</sup>. In some past cases, CMS or the United States Department of Justice have determined that certain behaviors are upcoding from whistleblower reports of fraud<sup>37,38</sup> or by manually auditing claims and electronic health records<sup>39,40</sup>, neither of which are scalable approaches. This is a substantial problem, as most major MA insurers have been accused of fraud related to their billing practices by whistleblowers or the United States Department of Justice<sup>38,39,41</sup>.

### *1.3. Contributions*

We expand RMTL estimation approaches for use in an impactful but understudied health policy application: evaluating incident HCC coding in MA following CMS-HCC formula updates. Here, individual HCCs and sets of mutually exclusive HCCs corresponding to severity levels of a single health condition are our events of interest, where an event occurs when an incident diagnosis is reported by an insurer. Besides being a unique application of TTE methods in itself, our approach expands on prior RMTL-based methodological work by proposing estimators for such analysis that can account for underreporting of events in TM, which, to our knowledge, has not been previously considered. Further, we propose a novel estimation approach for identifying one type of possible severity-based upcoding.

Our contributions also include the creation of an R package to simulate baseline HCC diagnoses as well as the ability to upcode or undercode these data. Our package simulates realistic co-occurring HCCs based on older American adults' self-reported health conditions from a large national survey. Self-reported data are not impacted by coding incentives to the same degree as billing claims or electronic health record data<sup>42–44</sup>, and enable analysis of incident coding without underreporting. Our package additionally allows users to modify these baseline data by either underreporting existing baseline diagnoses (realistic to TM data) or upcoding specific HCCs. Labeled upcoding data did not previously exist, so these simulated data are broadly useful for methods development in health policy.

## **2. Methods**

We outline our statistical approaches for TTE estimation in this section. In our setting, TTE and time to reported event are equivalent because we do not observe each coding event directly. However, since our estimates are over an entire monitoring period, we do not consider potential delays in reporting events to have notable impact. In addition, multiple events may be reported at each of several time points within a monitoring period (e.g., a monitoring period could be a calendar year where each time interval is three months). There also may be more than one monitoring period where reporting occurs. Given this, our goals are to both (1) evaluate reporting of incident events within one monitoring period and (2) compare incident event reporting across sequential monitoring periods. We focus on estimating differences in time to incident coding behaviors, as absolute measures are especially useful in health policy contexts<sup>45</sup> and new behaviors following policy changes are as well.

## 2.1. Notation

### 2.1.1. Incident events

An event (reported coding of either a single HCC or a single member of a set of HCCs corresponding to severity levels of one health condition) is considered incident if it was not reported in prior monitoring periods. The reporting of an incident event is represented by a vector  $S$ , with  $s \geq 1$  possible mutually exclusive subtype events.  $S$  encodes which of these events occurs first, analogous to competing risks. These are referred to as competing events and examined in a event-specific manner. When  $s > 1$ , the possible values  $S$  can take are written  $s \in \{1, \dots, k\}$  in order of increasing severity. Incident reporting of these events is observed in a set of two independent groups labeled by  $g$ , where  $g = 0$  is a comparison group for  $g = 1$ , over  $m \geq 2$  pre-specified monitoring periods.

For a given monitoring period, group, and subtype event,  $T$  is the true time to incident reporting for that event only. In addition,  $C$  is the time to censoring not due to any competing event. We observe  $Y = \min(T, C)$  and know whether censoring or reporting of the event happened first, which is denoted by  $\Delta = I(T \leq C)$ . Our observed data are therefore of the form  $\{(Y_1, S_1\Delta_1), \dots, (Y_n, S_n\Delta_n)\}$  for the  $n$  total events reported within the monitoring period. It is possible that multiple incident events could be reported at the same time. Ordered discrete event reporting times are  $t_1 < t_2 < \dots < \tau$ , where  $\tau$  is the end time of the monitoring period. A single time of event reporting in a monitoring period is denoted  $t_i$ . Finally, when comparing across groups or monitoring periods we add a respective  $g$  or  $m$  subscript to the notation above. Sequential monitoring periods are written as  $m - 1$  and  $m$ .

### 2.1.2. Reference events

We use a set of reference events, or HCCs distinct from the HCCs examined for incident reporting, to estimate underreporting. In theory, reference HCCs reported in one monitoring period should continue to be reported at equal rates in subsequent monitoring periods. However, in practice some reference events (and events in  $g = 0$  more broadly) may be underreported. Reference events are denoted as a vector  $S^*$  of  $h$  distinct HCCs, which can take values  $s^* \in \{1, \dots, h\}$ . None of these reference events have competing events. The persistence of a given reference event  $s^*$  in monitoring period  $m$ , meaning the proportion of individuals coded with  $s^*$  in monitoring period  $m$  who were previously coded with  $s^*$  in monitoring period  $m - 1$ , is  $q_{s^*,m}$  in line with prior work<sup>36</sup>.

## 2.2. Target estimands

Before describing our estimands for incident event reporting, we first describe some key components. Within a group  $g$  and monitoring period  $m$ ,  $F(t) = P(T \leq t)$  is the cumulative distribution function of  $T$  for all events beginning in that monitoring period. The overall hazard  $\lambda(t) = P(T = t | T > t)$  can be defined in terms of  $F(t)$  as  $\lambda(t) = (F(t) - F(t - 1))/(1 - F(t - 1))$ . In addition,  $\bar{F}(t) = 1 - F(t) = P(T > t) = \prod_{t_i \leq t} (1 - \lambda(t_i))$ , which is equivalent to the overall survival function in standard TTE analyses<sup>9</sup>.

The event-specific hazard is analogous to a cause-specific hazard in TTE analyses with competing risks. For event  $s$  within group  $g$  and monitoring period  $m$ , the event-specific hazard at  $t_i$  is  $\lambda_s(t_i) = P(T = t_i, S = s | T \geq t_i)$ . Then, the corresponding event-specific cumulative incidence is  $F_s(t) = P(T \leq t, S = s) = \sum_{t_i \leq t} \bar{F}(t_i) \lambda_s(t_i) = \sum_{t_i \leq t} \theta(t_i)$ <sup>9</sup>. Thus, the

mean time without event  $s$  is  $\mu_s(\tau) = \sum_{t_i < \tau} (t_{i+1} - t_i) F_s(t_i)$ <sup>10</sup>. If we are comparing across sequential monitoring periods or groups, we write  $\mu_{s,m}(\tau)$  to specify the monitoring period  $m$  and  $\mu_{s,g}(\tau)$  to specify group  $g$ .

Right censoring is assumed to be non-informative except for when an individual is censored due to a competing event. When an individual is recorded as having an event  $s$  that has any competing events (e.g., if  $s > 1$ ), they leave the risk set to be coded for any other competing event. Both  $g = 1$  and  $g = 0$  groups are expected have events recorded at all equivalent reporting times within the monitoring period. We also impose that the overall sample is fixed across all monitoring periods.

### 2.2.1. Underreporting in comparison group

The  $s^* \geq 1$  reference events occur in  $g = 0$  only. So, in the comparison group  $g = 0$  only and two sequential monitoring periods  $m - 1$  and  $m$ , our estimand for the underreporting proportion  $\epsilon$  is one minus the average persistence of all reference events, or  $\epsilon = 1 - \frac{1}{h} \sum q_{s^*,m}$ . Multiple pairs of sequential monitoring periods are distinguished by adding a subscript,  $\epsilon_m$ , where  $m$  denotes the second monitoring period in a pair.

### 2.2.2. Difference in mean time without event across groups

We are first interested in estimating the difference in incident reporting for event  $s$  across the two groups within one monitoring period. So, our estimand is  $\psi = \mu_{s,g=1}(\tau) - \mu_{s,g=0}(\tau)$ . In a monitoring period where  $\epsilon$  is known, the estimand is modified to account for underreporting by shifting the event-specific cumulative incidence curve in the comparison group  $g = 0$  by  $\epsilon$ , or  $\mu_{s,g=0}^*(\tau) = \sum_{t_i < \tau} (t_{i+1} - t_i) (F_s(t_i) + \epsilon)$ . This is written as  $\psi^* = \mu_{s,g=1}(\tau) - \mu_{s,g=0}^*(\tau)$ .

Across the two groups in two sequential monitoring periods, the reported difference in mean time without event that does not account for underreporting is  $\psi_M = \psi_m - \psi_{m-1}$ .  $\psi_m$  and  $\psi_{m-1}$  are the same as  $\psi$  but with an additional subscript to specify the monitoring period. This estimand can also be adjusted to account for underreporting when  $\epsilon_m$  and  $\epsilon_{m-1}$  are known, becoming  $\psi_M^* = \psi_m^* - \psi_{m-1}^*$ .

### 2.2.3. Possible severity-based upcoding

Possible severity-based upcoding can also be estimated within a monitoring period. Here,  $s \geq 2$ , where the least severe subtype corresponds to  $s = 1$  and the most severe subtype to  $s = k$ . In one monitoring period, possible severity-based upcoding across groups is  $\omega = \omega_{g=1} - \omega_{g=0}$ , where each  $\omega_g = \mu_{s=k,g}(\tau) - \mu_{s=1,g}(\tau)$ . This is equivalent to comparing the difference in incident reporting of the most severe event versus the least severe event across groups. If such reporting is higher in  $g = 1$  compared with  $g = 0$  (e.g.,  $\omega > 0$ ), then severity-based upcoding may be occurring.

## 2.3. Estimators

In order to introduce the estimator of  $\mu_s(\tau)$ , we first describe the estimator for the event-specific hazard at  $t_i$ :  $\hat{\lambda}_s(t_i) = d_{s,i}/r_i$ , where  $d_{s,i}$  is the count of event  $s$  reported at  $t_i$  and  $r_i$  is the number of individuals at risk at  $t_i$ , meaning the number of individuals who have not yet been recorded as having  $s$  or any competing event by  $t_i$ .  $d_i = \sum d_{s,i}$  is the count of all competing events reported at  $t_i$ . We also have the event-specific cumulative incidence

estimator:  $\hat{F}_s(t) = \sum_{t_i \leq t} \hat{\theta}(t_i)$ , with  $\hat{\theta}(t_i) = \widehat{F}(t_i) \times \hat{\lambda}_s(t_i)$ , where  $\widehat{F}(t) = \prod_{t_i \leq t} (1 - \hat{\lambda}(t_i)) = \prod_{t_i \leq t} (1 - (d_i/r_i))$  is the Kaplan-Meier estimator for  $\overline{F}(t) = \prod_{t_i \leq t} (1 - \lambda(t_i))$ <sup>9,46</sup>.

The variance for  $\hat{F}_s(t)$  is  $\widehat{\text{var}}(\hat{F}_s(t)) = \sum_{t_l < t} \widehat{\text{var}}(\hat{\theta}(t_l)) + 2 \sum_{t_l < t} \sum_{t_l < t_i \leq t} \widehat{\text{cov}}(\hat{\theta}(t_l), \hat{\theta}(t_i))$ , and covariance is  $\widehat{\text{cov}}(\hat{F}_s(t), \hat{F}_s(u)) = \widehat{\text{var}}(\hat{F}_s(t)) + \sum_{t_l \leq t} \sum_{t_l < t_i \leq u} \widehat{\text{cov}}(\hat{\theta}(t_l), \hat{\theta}(t_i))$ . Here  $t_l$  indicates an event reporting time prior to  $t_i$  and  $u$  is an arbitrary time distinct from  $t$ .  $\hat{\theta}(t_i)$  has variance  $\widehat{\text{var}}(\hat{\theta}(t_i)) = (\hat{\theta}(t_i))^2 \times ((r_i - d_{si})/(d_{si}r_i) + \sum_{t_l < t_i} (d_l/(r_l(r_l - d_l))))$  and covariance  $\widehat{\text{cov}}(\hat{\theta}(t_i), \hat{\theta}(t_l)) = \hat{\theta}(t_i)\hat{\theta}(t_l)(-(1/r_i) + \sum_{t_l < t_i} (d_l/(r_l(r_l - d_l))))$ <sup>10</sup>, where  $d_l = \sum d_{s,l}$  is the reported count of events of any subtype at  $t_l$ ,  $d_{s,l}$  is analogous to  $d_{s,i}$ , and  $r_l$  is the count of beneficiaries at risk at  $t_l$ .

We can now define the estimator of  $\mu_s(\tau)$  based on prior literature<sup>10</sup>, as this is a component of the novel estimators we develop next:  $\hat{\mu}_s(\tau) = \sum_{t_i < \tau} (t_{i+1} - t_i) \hat{F}_s(t_i)$ . This estimator has variance  $\widehat{\text{var}}(\hat{\mu}_s(\tau)) = \sum_{t_i < \tau} (t_{i+1} - t_i)^2 \widehat{\text{var}}(\hat{F}_s(t_i)) + 2 \sum_{t_i < \tau} \sum_{t_l < t_i} (t_{i+1} - t_i)(t_{l+1} - t_l) \widehat{\text{cov}}(\hat{F}_s(t_i), \hat{F}_s(t_l))$ .

### 2.3.1. Underreporting in comparison group

For each reported reference event  $s^*$  in  $g = 0$  we first compute persistence, or the proportion of individuals reported as having  $s^*$  in monitoring period  $m$  who were also reported as having  $s^*$  in monitoring period  $m - 1$ . This is then averaged across all reference events to obtain  $\hat{\epsilon} = 1 - \frac{1}{h} \sum \hat{q}_{s^*,m}$ .

### 2.3.2. Difference in mean time without event across groups

We propose that the difference in mean time without an event,  $\psi$ , is estimated across groups as  $\hat{\psi} = \hat{\mu}_{s,g=1}(\tau) - \hat{\mu}_{s,g=0}(\tau)$  with variance  $\widehat{\text{var}}(\hat{\psi}) = \widehat{\text{var}}(\hat{\mu}_{s,g=1}(\tau)) + \widehat{\text{var}}(\hat{\mu}_{s,g=0}(\tau))$ , as groups are assumed to be independent. To adjust for underreporting,  $\hat{\mu}_{s,g=0}(\tau) = \sum_{t_i < \tau} (t_{i+1} - t_i) \hat{F}_s(t_i)$  is modified to  $\hat{\mu}_{s,g=0}(\tau) = \sum_{t_i < \tau} (t_{i+1} - t_i) (\hat{F}_s(t_i) + \epsilon)$  in the estimator for  $\psi$ , which does not change the variance. The underreporting-adjusted estimator is denoted  $\hat{\psi}^*$ .

Across sequential monitoring periods,  $\psi_M$  is estimated as  $\hat{\psi}_M = \hat{\psi}_m - \hat{\psi}_{m-1}$  with variance  $\widehat{\text{var}}(\hat{\psi}_M) = \widehat{\text{var}}(\hat{\psi}_m) + \widehat{\text{var}}(\hat{\psi}_{m-1})$ . Although the components of this estimator come from the same fixed population, the correlation between component estimates is assumed to be negligible because we are comparing differences in RMTL across independent groups estimated in disjoint time intervals. Therefore, covariance is zero. When an underreporting estimate is known for both monitoring periods, this estimator becomes  $\hat{\psi}_M^* = \hat{\psi}_m^* - \hat{\psi}_{m-1}^*$ , and variance remains unchanged.

### 2.3.3. Possible severity-based upcoding

Our estimator of possible severity-based upcoding across groups,  $\omega$ , is given by:  $\hat{\omega} = \hat{\omega}_{g=1} - \hat{\omega}_{g=0}$ , where each  $\hat{\omega}_g = \hat{\mu}_{s=k}(\tau) - \hat{\mu}_{s=1}(\tau)$  within one monitoring period. The variance is  $\widehat{\text{var}}(\hat{\omega}) = \widehat{\text{var}}(\hat{\omega}_{g=1}) + \widehat{\text{var}}(\hat{\omega}_{g=0})$ , where  $\widehat{\text{var}}(\hat{\omega}_g) = \widehat{\text{var}}(\hat{\mu}_{s=k}(\tau)) + \widehat{\text{var}}(\hat{\mu}_{s=1}(\tau)) - 2\widehat{\text{cov}}(\hat{\mu}_{s=k}(\tau), \hat{\mu}_{s=1}(\tau))$  and  $\widehat{\text{cov}}(\hat{\mu}_{s=k}(\tau), \hat{\mu}_{s=1}(\tau)) = \sum_i \sum_j \widehat{\text{cov}}(\hat{F}_k(t_{i-1}), \hat{F}_1(t_{j-1}))$  for distinct event times  $t_i$  and  $t_j$ . Here,  $t_i$  corresponds to event reporting increments in the cumulative incidence function for severity level  $k$ , while  $t_j$  corresponds to the equivalent for severity level 1. Although event reporting times are the same, we write these using separate variables because covariance is estimated between all pairs of increments for these two severity levels.

### **3. upcoding R package**

#### *3.1. Challenges in identifying and estimating upcoding*

Upcoding in Medicare has been inconsistently examined in the medical and economics literature for several decades<sup>31</sup>. One reason for this is that it is difficult to definitively identify upcoding, particularly given that researchers and policymakers typically only have access to national Medicare claims data without beneficiaries' corresponding electronic health records or other data. There are also a number of barriers to the development of upcoding estimation approaches. First, gaining access to individual-level Medicare data is a time-consuming process that is inaccessible to many researchers. Second, there are limited national data resources describing co-occurring health conditions in older adults that are free of coding incentives<sup>42–44</sup>. Third, labeled upcoding data to evaluate estimators are not available to researchers. This also means that comparing proposed methods is challenging, as the data used to develop or evaluate such methods are often not able to be shared by researchers.

#### *3.2. Package functionality*

We developed the open source `upcoding` R package (<https://github.com/StanfordHPDS/upcoding>) to help address many of these issues. The package enables simulation of longitudinal coding data for a Medicare-eligible population as well as more reproducible evaluation of approaches for evaluating HCC coding. Features include:

- 1. Simulating a sample of individuals with realistic baseline HCCs.** These co-occurring baseline diagnoses are both based on older Americans' self-reported health conditions and free of upcoding and undercoding.
- 2. Upcoding baseline data to a specified level over multiple time points.** Users have the option to upcode any HCC using either any-available or severity-based upcoding. This results in labeled upcoding data, which is a useful resource for many types of coding measurement and estimation. In addition, users have the option to vary loss to follow up at each time point, a common issue in billing claims data and cohorts of older adults<sup>47</sup>.
- 3. Undercoding baseline data to a specified level.** Users have the option to specify an undercoding proportion, and that proportion of all existing diagnoses are removed from the overall dataset. This can help simulate data that is similar to TM, which may be of interest as a comparison group for analyses as well as the non-undercoded baseline data.

We describe this functionality in further detail in the subsections below. A brief tutorial on specific package functions is available in the package's Github repository.

##### *3.2.1. Baseline data simulation*

To simulate realistic co-occurring HCCs not influenced by coding incentives, co-occurring self-reported health conditions from participants aged 65 years and older (i.e., Medicare eligible) were extracted from the National Institutes of Health's All of Us study<sup>48</sup>. We used self-reported survey data to obtain co-occurring health conditions because other national datasets (e.g., Medicare billing claims) report diagnoses for billing purposes and therefore may be impacted by coding incentives<sup>42–44</sup>. The All of Us study was designed to enroll

a million participants across the United States, focusing especially on groups historically underrepresented in clinical and biomedical research<sup>48</sup> and included questions that overlapped with many HCCs in the current version of CMS-HCC (Version 28, or V28).

V28 HCCs (listed in the Supporting Information) were manually mapped to Systemized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) concept identifiers for survey questions (using version 7 of All of Us data; see this project's Github repository for mapping). Then, survey responses to all available SNOMED-CT concepts were queried from the All of Us database. Included surveys, available respondent sociodemographic characteristics, and coverage of V28 HCCs are described in the Supporting Information.

For each surveyed person, self-reported co-occurring V28 HCCs were extracted. This was then summarized into a table of unique co-occurring HCC sets and a count of the number of All of Us survey respondents that reported having these diagnoses. In line with the All of Us Data and Statistics Dissemination Policy, only sets of co-occurring HCCs with more than 21 respondents were exported from All of Us's platform (the Researcher Workbench) to be used in analyses. This both omits all sets of 20 or fewer respondents and also excludes several sets of co-occurring HCCs with more than 20 respondents. Ultimately, baseline data were simulated by sampling with replacement from these sets of co-occurring HCCs, where the respondent count is used to weight sampling. Users also have the option to use alternate sets of co-occurring V28 HCCs for baseline data sampling if they prefer.

### *3.2.2. Upcoding and loss to follow up simulation*

The package enables users to simulate two types of upcoding realistic to MA: any-available or severity-based, although the latter can only occur if an HCC has competing events. This upcoding is randomly split across a user-specified number of time points. Even though there is inherent right censoring in upcoded data (because only a proportion of all simulated individuals will be upcoded or coded at all), the package provides users with the option to include additional right censoring. This is meant to be representative of loss to follow up or death, both of which are known issues in following a population aged 65 years and older over time<sup>47</sup>. Users specify the proportion of loss to follow up they want to include, and a randomly selected set of rows (e.g., individuals) are right censored over each time point. Once someone is lost to follow up, they cannot be coded for any HCC in subsequent time periods.

### *3.2.3. Undercoding simulation*

The package also separately enables undercoding. Here, users specify a proportion of all coded diagnoses (from simulated baseline data) to randomly remove across the entire data set. This occurs on a dataset-wide level because undercoding is a systemic issue in TM<sup>28,33</sup> and so we do not assume that it impacts specific V28 HCCs disproportionately.

## **4. Simulation study**

Our simulation study demonstrates how the estimators we propose can be used to monitor Medicare coding behaviors. To do this, different upcoding and underreporting scenarios for the estimator introduced in Section 2.3.2 were compared to an estimator currently used by policymakers, defined later in Section 4.1. Degrees of upcoding and undercoding were

constructed to align with current estimates of these issues in MA and TM from the literature. In addition, the temporal structure was implemented to be analogous to quarterly reporting for the length of time (around two years) that a given risk adjustment formula version is typically in place. Each individual scenario was replicated 1000 times.

In each replicate, two sets of 1,000,000 observations of baseline data were independently simulated using our `upcoding` R package. For each dataset, the columns are the V28 HCCs (listed in the Supporting Information). Both any-available and severity-based upcoding were implemented in the first baseline data set (the MA-like data). For the former type of upcoding, an HCC without any competing HCCs was upcoded. As an illustration, we use HCC238 (Specified Heart Arrhythmias). For the latter, an HCC that has lower severity HCCs was upcoded, specifically HCC125 (Dementia, Severe). HCC125's less severe HCCs are HCC126 (Dementia, Moderate) and HCC127 (Dementia, Mild or Unspecified). The scenarios implemented were as follows:

**Scenario 1: Upcoding of an HCC that lacks competing events (HCC238) with varying underreporting in the comparison group.** Any observation not previously coded with HCC238 was eligible to be upcoded. Baseline MA data were separately upcoded to varying degrees (20%, 25%, 30%) sequentially within each monitoring period. The comparison group (i.e., TM data) was simulated by first undercoding the baseline TM data to varying levels (0%, 5%, 10%, 15%) and then upcoding HCC238 analogously but to a lower amount (5%) per monitoring period.

**Scenario 2: Upcoding of an HCC with lower-severity competing events (HCC125) with varying underreporting in the comparison group.** Only observations previously coded with the lower severity HCCs (HCC126 or HCC127) were eligible to be upcoded. Baseline MA data were separately upcoded to varying degrees (20%, 25%, 30%) sequentially within each monitoring period. The comparison group was simulated by first undercoding the baseline TM data to varying levels (0%, 5%, 10%, 15%) and then upcoding HCC125 analogously but to a lower amount (5%) per monitoring period.

#### 4.1. Comparator estimator

We compared our estimators to a coding intensity estimator similar to the Demographic Estimate of Coding Intensity (DECI), which is widely used by Medicare policymakers<sup>34,49</sup>. DECI assumes complete data (e.g., no censoring) and has the following formula:

$$\text{DECI} = \frac{\frac{\text{National average MA CMS-HCC risk score}}{\text{National average TM CMS-HCC risk score}}}{\frac{\text{National average MA demographic-only CMS-HCC risk score}}{\text{National average TM demographic-only CMS-HCC risk score}}}.$$

Here, the numerator is a risk score estimated with all HCCs included in the CMS-HCC formula, while the demographic-only risk adjustment risk score is estimated using only the demographic variables in that formula. Importantly, this also means that the estimand targeted by DECI is different from that of our proposed estimators. Further, as we did not have CMS-HCC risk score coefficients or demographic variables available in our simulated

data, our comparator estimator was more precisely DECI-like and we defined it as:

$$\text{DECI}^\dagger = \frac{\text{Average count of MA CMS-HCC HCCs}}{\text{Average count of TM CMS-HCC HCCs}}.$$

Our  $\text{DECI}^\dagger$  comparator is estimated in the same simulated data described earlier in this section, but relies on counts of all HCCs at the end of each monitoring period rather than time to incident coding of individual HCCs. Although DECI has a different target estimand than our estimators and ignores censoring, comparing our simulation results to the DECI-like estimator  $\text{DECI}^\dagger$  is useful as it illustrates how our estimators can be a complement to existing practice.

## 4.2. Simulation results

### 4.2.1. Proposed estimators

Cumulative incidence of coding for scenario 1 given 20% any-available upcoding in MA and 5% any-available upcoding in TM after all four degrees of undercoding is shown in Figure 1. As expected, within each monitoring period the cumulative incidence of the upcoded group was higher than the comparison group, which was only upcoded 5%. We also saw that the impact of undercoding on cumulative incidence estimates was limited. Given that undercoding occurs across the entire set of 115 HCCs, it was less likely to notably influence estimates for any single HCC. Lastly, we observed that the gaps between upcoded and comparison groups become smaller in sequential monitoring periods, which is a consequence of the fixed sample we imposed across all monitoring periods. As monitoring periods increased, the number of individuals available to upcode decreased. Results plots for additional degrees of upcoding (25%, 30%) and Scenario 2 upcoding (severity-based upcoding for HCC125) are in the Supporting Information.

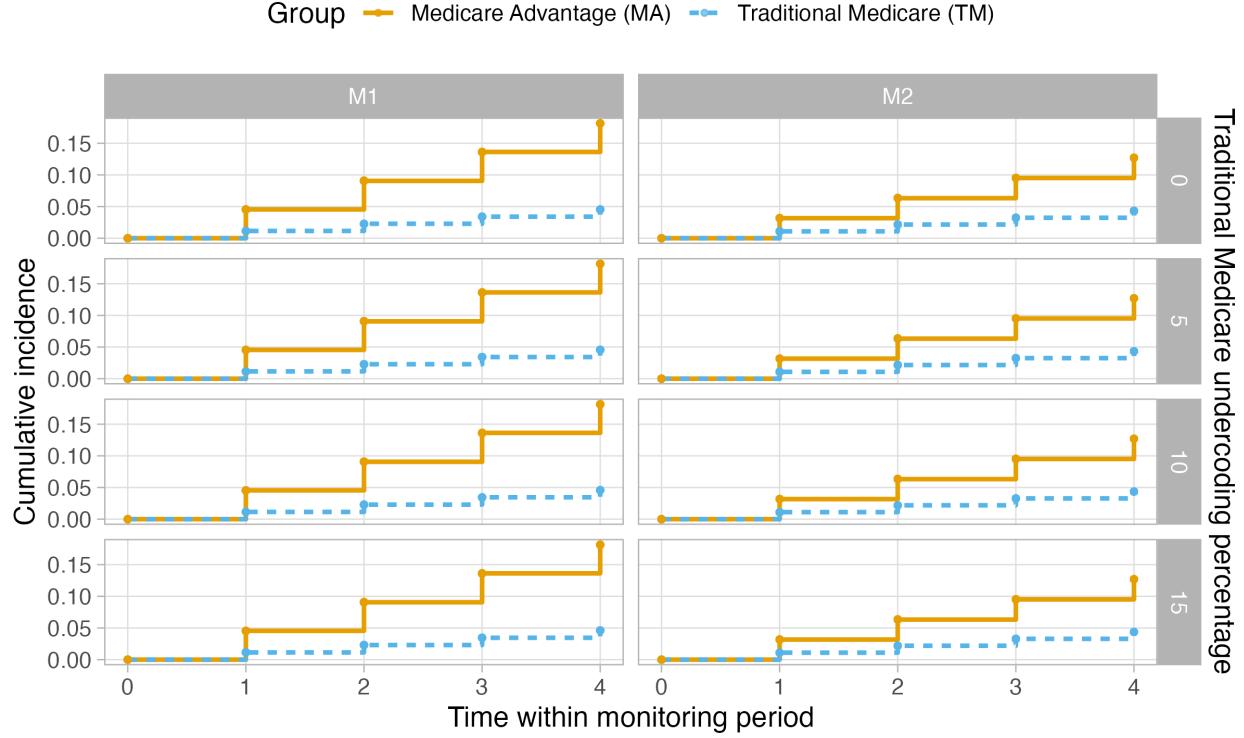


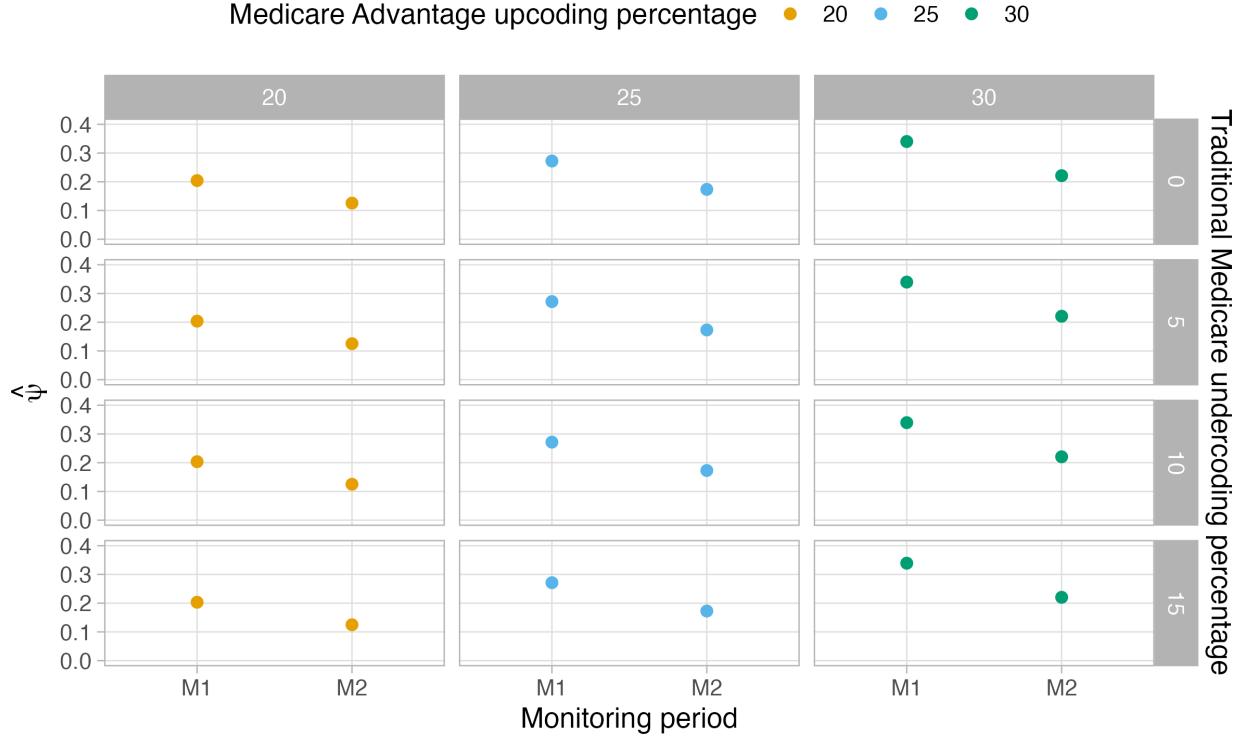
Figure 1: **Cumulative incidence functions for the Specified Heart Arrhythmias Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 20% any-available MA upcoding.** Specified heart arrhythmias corresponds to HCC238, which does not have any competing events. For this HCC, 20% of any-available individuals in the MA group are upcoded and 5% of any-available individuals in the TM comparison group are upcoded. The first monitoring period is labeled M1, and the second monitoring period is labeled M2. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

Estimates corresponding to the Section 2.3.2 estimators for upcoding of HCC238 are presented in Figure 2, which examines the difference in time without reported incident HCC238 coding between MA and TM within each monitoring period. Examining the first monitoring period (M1), estimates clearly increase with the degree of upcoding. Similarly, for monitoring period 2 (M2) alone, estimates also increase with the degree of upcoding.

In addition, undercoding has a limited effect on estimates. However, researchers also have the option to adjust for underreporting using the estimator proposed in Section 2.3.1. As the sample is fixed across monitoring periods, estimates across monitoring periods within a degree of upcoding and undercoding decrease, analogously to Figure 1. The gap between M1 and M2 increases with degree of upcoding as well, because more aggressive incident upcoding in M1 limits the availability of individuals for incident upcoding in M2.

This suggests that within a monitoring period, researchers could compare estimates of different HCCs to obtain a ranked list of HCCs to study further, where HCCs with the highest estimates have the most incident coding. In addition, in this scenario, being in TM appeared to have a “protective effect” against being coded with HCC238, although

supplemental analyses would be needed to account for possible confounding and other issues. Additional results plots for severity-based upcoding of HCC125 are available in the Supporting Information.



**Figure 2: Within-monitoring period period  $\psi$  estimates for the Specified Heart Arrhythmias Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups.** Specified heart arrhythmias corresponds to HCC238, which does not have any competing events. For this HCC, any-available individuals in the MA group are upcoded to varying degrees and any-available individuals in the TM comparison group are upcoded 5%. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

#### 4.2.2. Comparator estimator

Figure 3 shows the average DECI<sup>†</sup> estimate across simulation replicates by monitoring period as well as upcoding and undercoding degree. As degree of upcoding increases, these estimates increase negligibly both for individual monitoring periods and across sequential monitoring periods. This is intuitive—only two of the 115 HCCs in the MA risk adjustment algorithm are being upcoded—but it suggests a limitation of DECI in identifying upcoding occurring in a minority of HCCs. Regardless of upcoding degree, the DECI<sup>†</sup> estimate increasing in line with the degree of TM undercoding indicates that it may be more sensitive to undercoding in TM. Especially since we know undercoding is prevalent in TM<sup>28,33</sup>, this suggests that DECI may be misrepresenting the overall spending gap between MA and TM due to its undercoding sensitivity.

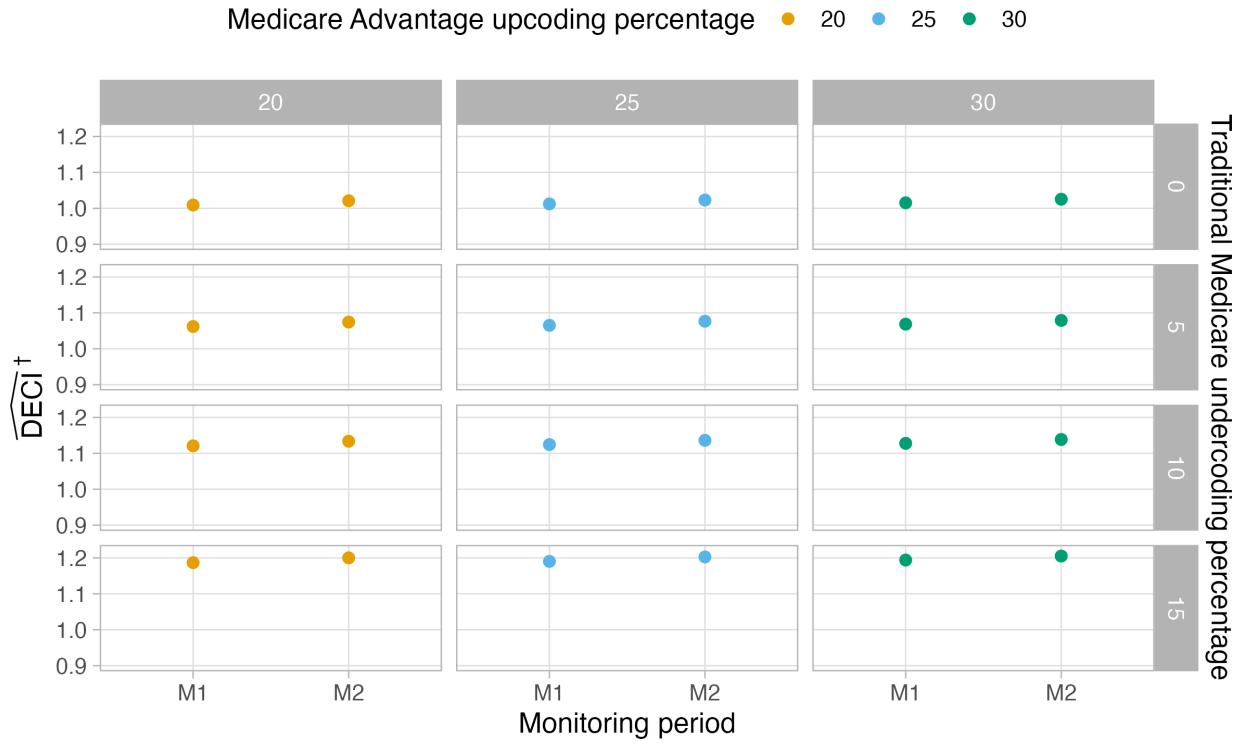


Figure 3: **DECI<sup>†</sup>** estimate across all Medicare Advantage (MA) Version 28 Hierarchical Condition Categories (HCCs) at varying degrees of upcoding and underreporting in simulated MA and Traditional Medicare (TM) groups. Three separate degrees of MA group upcoding occur sequentially over each monitoring period in HCC238 (any-available) and HCC125 (lower severity) only. The TM group is upcoded 5% sequentially for the same HCCs over equivalent periods. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

## 5. Discussion

We proposed a set of estimators that extend RMTL methodology for evaluating time to incident coding by private insurers in Medicare. Given the timeline of risk adjustment formula updates, our approach realistically presupposes that reported coding is examined over multiple monitoring periods, each of which has several time points where new events are reported. Novel estimators were introduced to evaluate differences in time-to-reporting both within monitoring periods and across monitoring periods, including an estimator for severity-based upcoding. Our approach also included an adjustment for possible underreporting, which is a known major issue in TM data<sup>28,33</sup>. Finally, we developed an open source R package that enables users to simulate co-occurring HCCs free of coding incentives and similar to those reported by individuals residing in the United States eligible for Medicare. Users can also undercode or upcode this data over time.

Our simulated data were upcoded to degrees aligned to those reported in literature<sup>32</sup>, and we found in our simulation results that our estimators were able to recover differences in upcoding both within and across monitoring periods while showing limited sensitivity to undercoding. DECI-like estimates of the same data were a useful complement, but were

limited in that these comparator estimates were very sensitive to undercoding and did not vary when upcoding occurred in a minority of HCCs. Therefore, our estimators show considerable promise as tools to help evaluate reported coding patterns over time. These estimators can be used as a first step to identify potentially upcoded HCCs while making more realistic assumptions than DECI. Follow up analyses could include examining coding patterns within specific insurers, providers, or beneficiaries.

This work has a number of limitations and areas for future development. First, several assumptions could be further relaxed. This includes our assumptions that the population is fixed across both comparison groups and monitoring periods and that estimates in sequential monitoring periods across groups are approximately independent. Upcoding estimators could be expanded to additional types of upcoding and to correct for underreporting. Covariates could also be incorporated into estimation, which would be useful for addressing issues like confounding. There is also more heterogeneity and missingness in real-world claims data than was included in our simulations. Self-reported diagnoses—as we use in our simulated baseline data—also have recall bias and other issues that we do not address here.

Since the MA program’s inception, high intensity of private insurer coding, including possible upcoding, is estimated to have cost the federal government and taxpayers \$224 billion dollars without clear benefit to beneficiaries<sup>32</sup>. Thus, estimating differences in coding between MA and TM or more and less severe competing HCCs has the potential to help policymakers locate issues with specific HCCs earlier and at scale. This could suggest areas for improvement in the risk adjustment formula and potentially save significant funds in the Medicare program. Our work has provided both novel estimators and novel simulated data tailored to addressing these important policy considerations.

## 6. Acknowledgments

This research was funded by National Institutes of Health Director’s Pioneer Award DP1LM014278 and a Stanford Interdisciplinary Graduate Fellowship. We thank Malcolm Barrett, Gabriela Basel, Lizzie Kumar, and Neha Srivathsra for their valuable insights and contributions to code performance and review. We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health’s All of Us Research Program (<https://allofus.nih.gov/>) for making available the participant survey data included in this study. We thank Stanford University and Stanford Research Computing for providing additional computational resources, including support for computing performed on the Sherlock cluster.

## 7. Data availability statement

The simulated data for this study can be generated using code in the project repository: [https://github.com/StanfordHPDS/tte\\_estimation\\_medicare](https://github.com/StanfordHPDS/tte_estimation_medicare). Additional upcoding, undercoding, and baseline data can be simulated using the upcoding package (<https://github.com/StanfordHPDS/upcoding>). In line with program policies, non-summarized version 7 All of Us survey data used to derive baseline co-occurring HCCs can be accessed by authorized users via the All of Us Researcher Workbench only (<https://www.researchallofus.org/data>).

[tools/workbench/](#)) and requires Registered Tier access. Code that was used for this project in the Researcher Workbench is in the above project repository.

## **8. Supporting Information**

Additional study summary information and results can be found online in the Supporting Information.

## References

1. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1):13-15.
2. Chappell R, Zhu X. Describing differences in survival curves. *JAMA Oncol*. 2016;2(7):906-907.
3. Uno H, Horiguchi M, Hassett MJ. Statistical test/estimation methods used in contemporary phase III cancer randomized controlled trials with time-to-event outcomes. *Oncologist*. 2020;25(2):91-93.
4. Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Med Res Methodol*. 2019;19(1):103.
5. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol*. 2016;34(15):1813-1819.
6. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*. 2014;15(2):222-233.
7. Tian L, Jin H, Uno H, et al. On the empirical choice of the time window for restricted mean survival time. *Biometrics*. 2020;76(4):1157-1166.
8. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: A review and recommendations for improvement. *Stat Med*. 2017;36(8):1203-1209.
9. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. 1st Edition. Chapman & Hall/CRC; 2020.
10. Conner SC, Trinquart L. Estimation and modeling of the restricted mean time lost in the presence of competing risks. *Stat Med*. 2021;40(9):2177-2196.
11. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg (Lond)*. 1949;47(2):188.
12. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409-2421.
13. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32(22):2380-2385.

14. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis: On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72(1):215-221.
15. Lu Y, Tian L. Statistical considerations for sequential analysis of the restricted mean survival time for randomized clinical trials. *Stat Biopharm Res*. 2021;13(2):210-218.
16. Zhang P, Logan B, Martens M. Covariate-adjusted group sequential comparisons of restricted mean survival times. *arXiv [statME]*. Published online September 2025.
17. Paukner M, Chappell R. Window mean survival time. *Stat Med*. 2021;40(25):5521-5533.
18. Sun J, Schaubel DE, Tchetgen EJT. Beyond fixed restriction time: Adaptive restricted mean survival time methods in clinical trials. *arXiv [statME]*. Published online January 2025.
19. Horiguchi M, Uno H. On permutation tests for comparing restricted mean survival time with small sample from randomized trials. *Stat Med*. 2020;39(20):2655-2670.
20. Ditzhaus M, Yu M, Xu J. Studentized permutation method for comparing two restricted mean survival times with small sample from randomized trials. *Stat Med*. 2023;42(13):2226-2240.
21. Munko M, Ditzhaus M, Dobler D, Genuneit J. RMST-based multiple contrast tests in general factorial designs. *Stat Med*. 2024;43(10):1849-1866.
22. Jesse D, Huber C, Friede T. Comparing restricted mean survival times in small sample clinical trials using pseudo-observations. *arXiv [statME]*. Published online August 2024.
23. Andersen PK. Decomposition of number of life years lost according to causes of death. *Stat Med*. 2013;32(30):5278-5285.
24. Calkins KL, Canan CE, Moore RD, Lesko CR, Lau B. An application of restricted mean survival time in a competing risks setting: Comparing time to ART initiation by injection drug use. *BMC Med Res Methodol*. 2018;18(1):27.
25. Wu H, Yuan H, Yang Z, Hou Y, Chen Z. Implementation of an alternative method for assessing competing risks: Restricted mean time lost. *Am J Epidemiol*. 2022;191(1):163-172.
26. Ochieng N, Cubanski J, Neuman T. A snapshot of sources of coverage among medicare beneficiaries. Published online December 2025.

27. Pope GC, Kautter J, Ellis RP, et al. Risk adjustment of medicare capitation payments using the CMS-HCC model. *Health Care Financ Rev*. 2004;25(4):119-141.
28. Ellis RP, Martins B, Rose S. Chapter 3 - risk adjustment for health plan payment. In: McGuire TG, Kleef RC van, eds. *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets*. Academic Press; 2018:55-104.
29. Centers for Medicare and Medicaid Services. *2024 MA Advance Notice.*; 2023.
30. Geruso M, Layton T. Upcoding: Evidence from Medicare on Squishy Risk Adjustment. *Journal of Political Economy*. 2020;128(3):984-1026. doi:[10.1086/704756](https://doi.org/10.1086/704756)
31. Joiner KA, Lin J, Pantano J. Upcoding in medicare: where does it matter most? *Health Economics Review*. 2024;14(1). doi:[10.1186/s13561-023-00465-4](https://doi.org/10.1186/s13561-023-00465-4)
32. Medicare Payment Advisory Commission. March 2025 report to the congress: Medicare payment policy. Published online March 2025.
33. Ghoshal-Datta N, Chernew ME, McWilliams JM. Lack of persistent coding in traditional medicare may widen the risk-score gap with medicare advantage. *Health Aff (Millwood)*. 2024;43(12):1638-1646.
34. Kronick R, Chua FM. Industry-Wide and Sponsor-Specific Estimates of Medicare Advantage Coding Intensity. *SSRN Electronic Journal*. Published online 2021. doi:[10.2139/ssrn.3959446](https://doi.org/10.2139/ssrn.3959446)
35. Jacobs PD, Kronick R. Getting What We Pay For: How Do Risk-Based Payments to Medicare Advantage Plans Compare with Alternative Measures of Beneficiary Health Risk? *Health Services Research*. 2018;53(6):4997-5015. doi:[10.1111/1475-6773.12977](https://doi.org/10.1111/1475-6773.12977)
36. McGuire TG, Enache OM, Chernew M, McWilliams JM, Nham T, Rose S. Incidence, persistence, and steady-state prevalence in coding intensity for health plan payment. *Health Serv Res*. 2025;(e70065):e70065.
37. Abelson R, Sanger-Katz M. “The cash monster was insatiable”: How insurers exploited medicare for billions. *The New York Times*. Published online October 2022.
38. Justice Office of Public Affairs USD of. Kaiser permanente affiliates pay \$556M to resolve false claims act allegations. Published online January 2026.
39. Abelson R, Sanger-Katz M. New medicare rule aims to take back \$4.7 billion from insurers. *The New York Times*. Published online January 2023.

40. Centers for Medicare & Medicaid Services. Medicare and medicaid programs; policy and technical changes to the medicare advantage, medicare prescription drug benefit, program of all-inclusive care for the elderly (PACE), medicaid fee-for-service, and medicaid managed care programs for years 2020 and 2021. *Federal Register*. 2023;88:6643-6665.
41. Constantino AK. UnitedHealth says it is cooperating with DOJ investigations into medicare billing practices. Published online July 2025.
42. Rose S, Zaslavsky AM, McWilliams JM. Variation In Accountable Care Organization Spending And Sensitivity To Risk Adjustment: Implications For Benchmarking. *Health Affairs*. 2016;35(3):440-448. doi:[10.1377/hlthaff.2015.1026](https://doi.org/10.1377/hlthaff.2015.1026)
43. Chernew ME, Carichner J, Impreso J, et al. Coding-Driven Changes In Measured Risk In Accountable Care Organizations. *Health Affairs*. 2021;40(12):1909-1917. doi:[10.1377/hlthaff.2021.00361](https://doi.org/10.1377/hlthaff.2021.00361)
44. McWilliams JM, Weinreb G, Landrum MB, Chernew ME. Use of patient health survey data for risk adjustment to limit distortionary coding incentives in medicare: Article examines use of patient health survey data for risk adjustment to limit distortionary coding incentives in medicare. *Health Aff (Millwood)*. 2025;44(1):48-57.
45. Westreich D. *Epidemiology by Design*. Oxford University PressNew York; 2019. doi:[10.1093/oso/9780190665760.001.0001](https://doi.org/10.1093/oso/9780190665760.001.0001)
46. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
47. Barberio J, Naimi AI, Patzer RE, et al. Influence of incomplete death information on cumulative risk estimates in US claims data. *Am J Epidemiol*. 2024;193(9):1281-1290.
48. The All of Us Research Program Investigators. The “All of Us” Research Program. *New England Journal of Medicine*. 2019;381(7):668-676. doi:[10.1056/nejmsr1809937](https://doi.org/10.1056/nejmsr1809937)
49. Medicare Payment Advisory Commission. *Report to Congress: Medicare Payment Policy*; 2024.

# Supporting Information

## Time-to-Event Estimation with Unreliably Reported Events in Medicare Health Plan Payment

Oana M. Enache<sup>a,\*</sup>, Sherri Rose<sup>b</sup>

<sup>a</sup>*Stanford University School of Medicine, Department of Biomedical Data Science, Edwards Building, 300 Pasteur Drive, Stanford, CA, 94304*

<sup>b</sup>*Stanford University, Department of Health Policy, Encina Commons, 615 Crothers Way, Stanford, CA, 94305*

---

---

### 1. Additional information about simulated data

Table 1: **Hierarchical Condition Categories (HCCs) included in the Medicare Advantage risk adjustment algorithm version 28 (V28).** Each of the HCCs in this table are described as written in the Centers for Medicare and Medicaid Services' (CMS) 2024 Medicare Advantage Advance Notice<sup>1</sup>. Competing events, if applicable, are also noted from the 2024 CMS Risk Adjustment Model Software and ICD-10 Mappings, specifically the Midyear/Final Model Software, which is in the SAS language<sup>2</sup>. If a set of competing events is not N/A, then, if the HCC in that row is coded, none of the competing event HCCs can also be coded for billing purposes. The hierarchy described in the “Competing Events” column can also be found in the `upcoding` R package.

HCC	HCC Description	Competing Events
HCC1	HIV/AIDS	N/A
HCC2	Septicemia, sepsis, systemic inflammatory response syndrome/shock	N/A
HCC6	Opportunistic infections	N/A
HCC17	Cancer metastatic to lung, liver, brain, and other organs; acute myeloid leukemia except promyelocytic	HCC18, HCC19, HCC20, HCC21, HCC22, HCC23
HCC18	Cancer metastatic to bone, other and unspecified metastatic cancer; acute leukemia except myeloid	HCC19, HCC20, HCC21, HCC22, HCC23

---

\*Corresponding author

Email address: oenache@stanford.edu (Oana M. Enache)

HCC	HCC Description	Competing Events
HCC19	Myelodysplastic syndromes, multiple myeloma, and other cancers	HCC20, HCC21, HCC22, HCC23
HCC20	Lung and other severe cancers	HCC21, HCC22, HCC23
HCC21	Lymphoma and other cancers	HCC22, HCC23
HCC22	Bladder, colorectal, and other cancers	HCC23
HCC23	Prostate, breast, and other cancers and tumors	N/A
HCC35	Pancreas transplant status	HCC36, HCC37, HCC38
HCC36	Diabetes with severe acute complications	HCC37, HCC38
HCC37	Diabetes with chronic complications	HCC38
HCC38	Diabetes with glycemic, unspecified, or no complications	N/A
HCC48	Morbid obesity	N/A
HCC49	Specified lysosomal storage disorders	N/A
HCC50	Amyloidosis, porphyria, and other specified metabolic disorders	N/A
HCC51	Addison's and Cushing's diseases, acromegaly, and other specified endocrine disorders	N/A
HCC62	Liver transplant status/complications	HCC63, HCC64, HCC65, HCC68
HCC63	Chronic liver failure/end-stage liver disorders	HCC64, HCC65, HCC68, HCC202
HCC64	Cirrhosis of liver	HCC65, HCC68
HCC65	Chronic hepatitis	N/A
HCC68	Cholangitis and obstruction of bile duct without gallstones	N/A
HCC77	Intestine transplant status/complications	HCC78, HCC80, HCC81
HCC78	Intestinal obstruction/perforation	N/A
HCC79	Chronic pancreatitis	N/A

HCC	HCC Description	Competing Events
HCC80	Crohn's disease (regional enteritis)	HCC81
HCC81	Ulcerative colitis	N/A
HCC92	Bone/joint/muscle/severe soft tissue infections/necrosis	N/A
HCC93	Rheumatoid arthritis and other specified inflammatory rheumatic disorders	HCC94
HCC94	Systemic lupus erythematosus and other specified systemic connective tissue disorders	N/A
HCC107	Sickle cell anemia (Hb-SS) and thalassemia beta zero	HCC108
HCC108	Sickle cell disorders, except sickle cell anemia (Hb-SS) and thalassemia beta zero; beta thalassemia major	N/A
HCC109	Acquired hemolytic, aplastic, and sideroblastic anemias	N/A
HCC111	Hemophilia, male	HCC112
HCC112	Immune thrombocytopenia and specified coagulation defects and hemorrhagic conditions	N/A
HCC114	Common variable and combined immunodeficiencies	HCC115
HCC115	Specified immunodeficiencies and white blood cell disorders	N/A
HCC125	Dementia, severe	HCC126, HCC127
HCC126	Dementia, moderate	HCC127
HCC127	Dementia, mild or unspecified	N/A
HCC135	Drug use with psychotic complications	HCC136, HCC137, HCC138, HCC139
HCC136	Alcohol use with psychotic complications	HCC137, HCC138, HCC139
HCC137	Drug use disorder, moderate/severe, or drug use with non-psychotic complications	HCC138, HCC139

HCC	HCC Description	Competing Events
HCC138	Drug use disorder, mild, uncomplicated, except cannabis	HCC139
HCC139	Alcohol use disorder, moderate/severe, or alcohol use with specified non-psychotic complications	N/A
HCC151	Schizophrenia	HCC152, HCC153, HCC154, HCC155
HCC152	Psychosis, except schizophrenia	HCC153, HCC154, HCC155
HCC153	Personality disorders; anorexia/bulimia nervosa	HCC154, HCC155
HCC154	Bipolar disorders without psychosis	HCC155
HCC155	Major depression, moderate or severe, without psychosis	N/A
HCC180	Quadriplegia	HCC181, HCC182, HCC253, HCC254
HCC181	Paraplegia	HCC182, HCC254
HCC182	Spinal cord disorders/injuries	N/A
HCC190	Amyotrophic lateral sclerosis and other motor neuron disease, spinal muscular atrophy	N/A
HCC191	Quadriplegic cerebral palsy	HCC180, HCC181, HCC182, HCC192, HCC253, HCC254
HCC192	Cerebral palsy, except quadriplegic	HCC180, HCC181, HCC182, HCC253, HCC254
HCC193	Chronic inflammatory demyelinating polyneuritis and multifocal motor neuropathy	N/A
HCC195	Myasthenia gravis with (acute) exacerbation	HCC196
HCC196	Myasthenia gravis without (acute) exacerbation and other myoneural disorders	N/A
HCC197	Muscular dystrophy	N/A
HCC198	Multiple sclerosis	N/A
HCC199	Parkinson and other degenerative disease of basal ganglia	N/A

HCC	HCC Description	Competing Events
HCC200	Friedreich and other hereditary ataxias; Huntington disease	N/A
HCC201	Seizure disorders and convulsions	N/A
HCC202	Coma, brain compression/anoxic damage	N/A
HCC211	Respirator dependence/tracheostomy status/complications	HCC212, HCC213
HCC212	Respiratory arrest	HCC213
HCC213	Cardio-respiratory failure and shock	N/A
HCC221	Heart transplant status/complications	HCC222, HCC223, HCC224, HCC225, HCC226, HCC227
HCC222	End stage heart failure	HCC223, HCC224, HCC225, HCC226, HCC227
HCC223	Heart failure with heart assist device/artificial heart	HCC224, HCC225, HCC226, HCC227
HCC224	Acute on chronic heart failure	HCC225, HCC226, HCC227
HCC225	Acute heart failure (excludes acute on chronic)	HCC226, HCC227
HCC226	Heart failure, except end stage and acute	HCC227
HCC227	Cardiomyopathy/myocarditis	N/A
HCC228	Acute myocardial infarction	HCC229
HCC229	Unstable angina and other acute ischemic heart disease	N/A
HCC238	Specified heart arrhythmias	N/A
HCC248	Intracranial hemorrhage	HCC249
HCC249	Ischemic or unspecified stroke	N/A
HCC253	Hemiplegia/hemiparesis	HCC254
HCC254	Monoplegia, other paralytic syndromes	N/A
HCC263	Atherosclerosis of arteries of the extremities with ulceration or gangrene	HCC264, HCC383, HCC409
HCC264	Vascular disease with complications	N/A
HCC267	Deep vein thrombosis and pulmonary embolism	N/A

HCC	HCC Description	Competing Events
HCC276	Lung transplant status/complications	HCC277, HCC278, HCC279, HCC280
HCC277	Cystic fibrosis	HCC278, HCC279, HCC280
HCC278	Idiopathic pulmonary fibrosis and lung involvement in systemic sclerosis	HCC279, HCC280
HCC279	Severe persistent asthma	HCC280
HCC280	Chronic obstructive pulmonary disease, interstitial lung disorders, and other chronic lung disorders	N/A
HCC282	Aspiration and specified bacterial pneumonias	HCC283
HCC283	Empyema, lung abscess	N/A
HCC298	Severe diabetic eye disease, retinal vein occlusion, and vitreous hemorrhage	N/A
HCC300	Exudative macular degeneration	N/A
HCC326	Chronic kidney disease, stage 5	HCC327, HCC328, HCC329
HCC327	Chronic kidney disease, severe (stage 4)	HCC328, HCC329
HCC328	Chronic kidney disease, moderate (stage 3B)	HCC329
HCC329	Chronic kidney disease, moderate (stage 3, except 3B)	N/A
HCC379	Pressure ulcer of skin with necrosis through to muscle, tendon, or bone	HCC380, HCC381, HCC382, HCC383
HCC380	Chronic ulcer of skin, except pressure, through to bone or muscle	HCC381, HCC382, HCC383
HCC381	Pressure ulcer of skin with full thickness skin loss	HCC382, HCC383
HCC382	Pressure ulcer of skin with partial thickness skin loss	HCC383
HCC383	Chronic ulcer of skin, except pressure, not specified as through to bone or muscle	N/A
HCC385	Severe skin burn	N/A

HCC	HCC Description	Competing Events
HCC387	Pemphigus, pemphigoid, and other specified autoimmune skin disorders	N/A
HCC397	Major head injury with loss of consciousness (> 1 hour)	HCC202, HCC398, HCC399
HCC398	Major head injury with loss of consciousness (< 1 hour or unspecified)	HCC202, HCC399
HCC399	Major head injury without loss of consciousness	N/A
HCC401	Vertebral fractures without spinal cord injury	N/A
HCC402	Hip fracture/dislocation	N/A
HCC405	Traumatic amputations and complications	HCC409
HCC409	Amputation status, lower limb/amputation complications	N/A
HCC454	Stem cell, including bone marrow, transplant status/complications	N/A
HCC463	Artificial openings for feeding or elimination	N/A

### 1.1. Summary statistics from All of Us survey data

These data represent included survey respondents with at least one self-reported health condition that mapped to a Hierarchical Condition Category (HCC) in Version 28 of the Medicare Advantage risk adjustment algorithm. Only sets of co-occurring HCCs with more than 21 respondents (which also excludes several sets with more than 20 respondents) were used in any summary tables to comply with the All of Us Data and Statistics Dissemination Policy.

Table 2: **Included surveys from All of Us**

All of Us Survey Title	Number of Unique Respondents
Overall Health	156
Personal and Family Health History	15692

Table 3: **Number of respondents by HCC with available survey questions**

Version 28 HCC	Number of Respondents
HCC1	546
HCC20	339
HCC21	5775
HCC22	260
HCC23	6772
HCC35	156
HCC38	2114
HCC51	4283
HCC62	156
HCC64	100
HCC65	546
HCC77	156
HCC78	511
HCC93	1131
HCC109	1018
HCC155	1467
HCC182	47
HCC221	156
HCC226	73
HCC228	399
HCC238	1635
HCC249	373
HCC264	62
HCC267	123
HCC276	156
HCC280	273
HCC300	587
HCC327	31
HCC328	110
HCC398	808

Table 4: **Distribution of HCC-relevant conditions per person**

Number of HCC-Relevant Conditions Per Person	Number of Respondents
1	5636
2	6657
3	3163
4	236
5	156

The following tables show respondent information using categories provided by All of Us. Not all respondents responded to all questions so counts may vary. The rows in each table correspond to available fields in All of Us. These characteristics are not displayed jointly because that would result in sample sizes too small to comply with the All of Us Data and Statistics Dissemination Policy.

Table 5: **Respondent age at survey response**

Age Group (years)	Number of Respondents	Percent of Respondents (%)
[65-70)	5088	32
[70-75)	5767	36
[75-80)	3383	21
[80-85)	1362	9
[85-90)	250	2

Table 6: **Respondents' self-reported sex at birth**

Self-Reported Sex at Birth	Number of Respondents	Percent of Respondents (%)
Female	8461	53
Male	6975	44
No matching concept	269	2
Not male, not female, prefer not to answer, or skipped	143	1

Table 7: **Respondents' self-reported ethnicity**

Self-Reported Ethnicity	Number of Respondents	Percent of Respondents (%)
Hispanic or Latino	633	4
Not Hispanic or Latino	14442	91
Prefer not to answer	34	0
Skipped	639	4
Neither	100	1

Table 8: **Respondents' self-reported race**

Self-Reported Race	Number of Respondents	Percent of Respondents (%)
Asian	260	2
Black or African American	752	5
White	13347	84
Another single population	49	0

Self-Reported Race	Number of Respondents	Percent of Respondents (%)
More than one population	129	1
I prefer not to answer	34	0
Skipped	639	4
None indicated	538	3
None of these	100	1

## 2. Additional simulation results

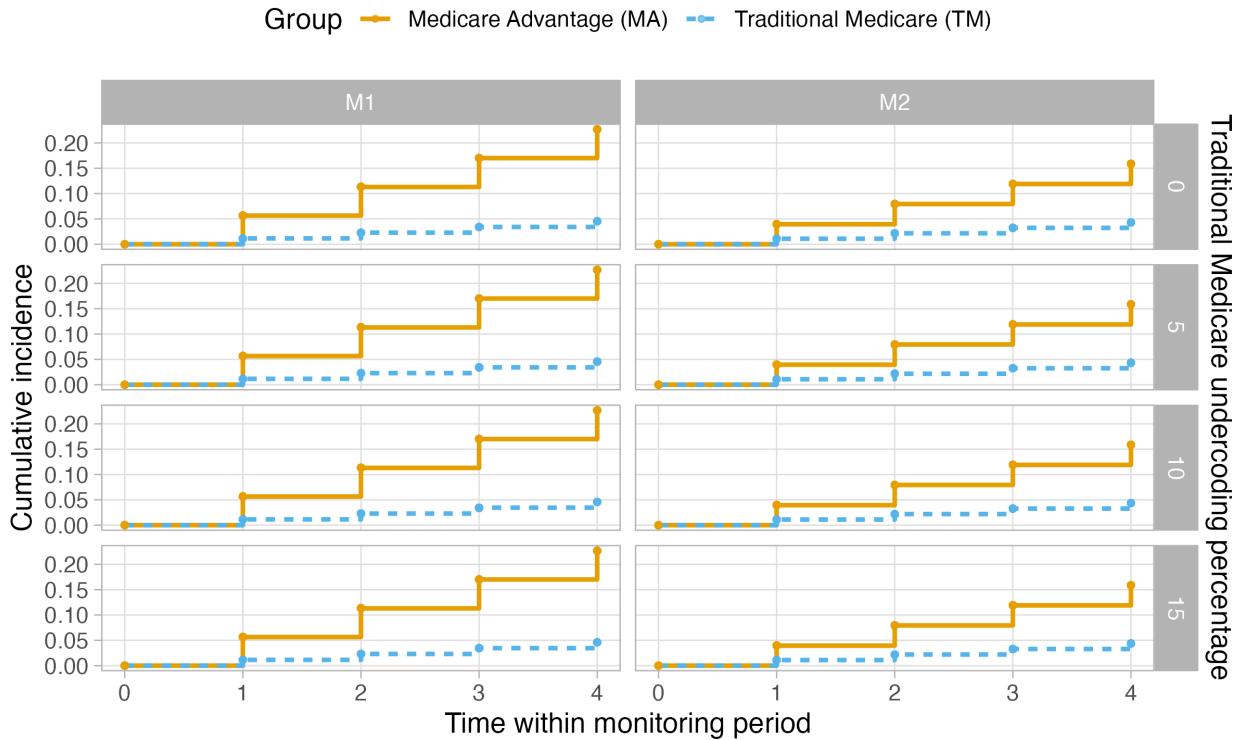


Figure 1: Cumulative incidence functions for the Specified Heart Arrhythmias Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 25% any-available MA upcoding. Specified heart arrhythmias corresponds to HCC238, which does not have any competing events. For this HCC, 25% of any-available individuals in the MA group are upcoded and 5% of any-available individuals in the TM comparison group are upcoded. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

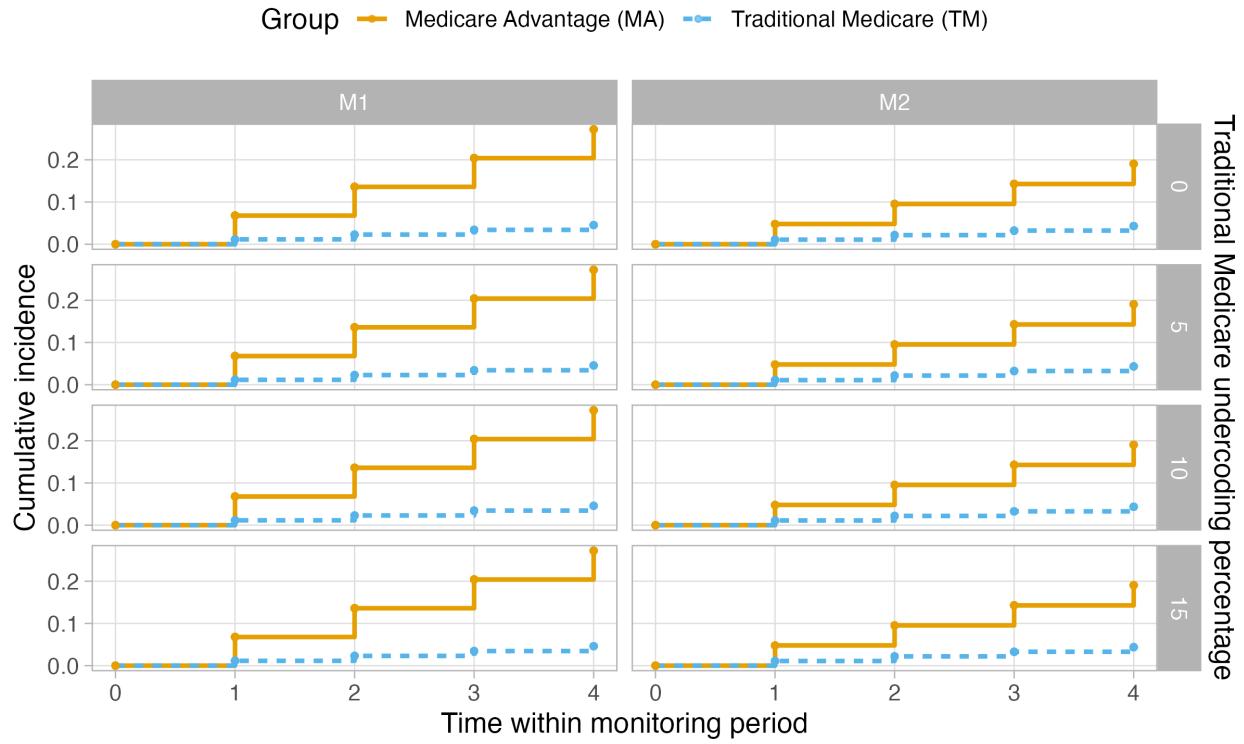
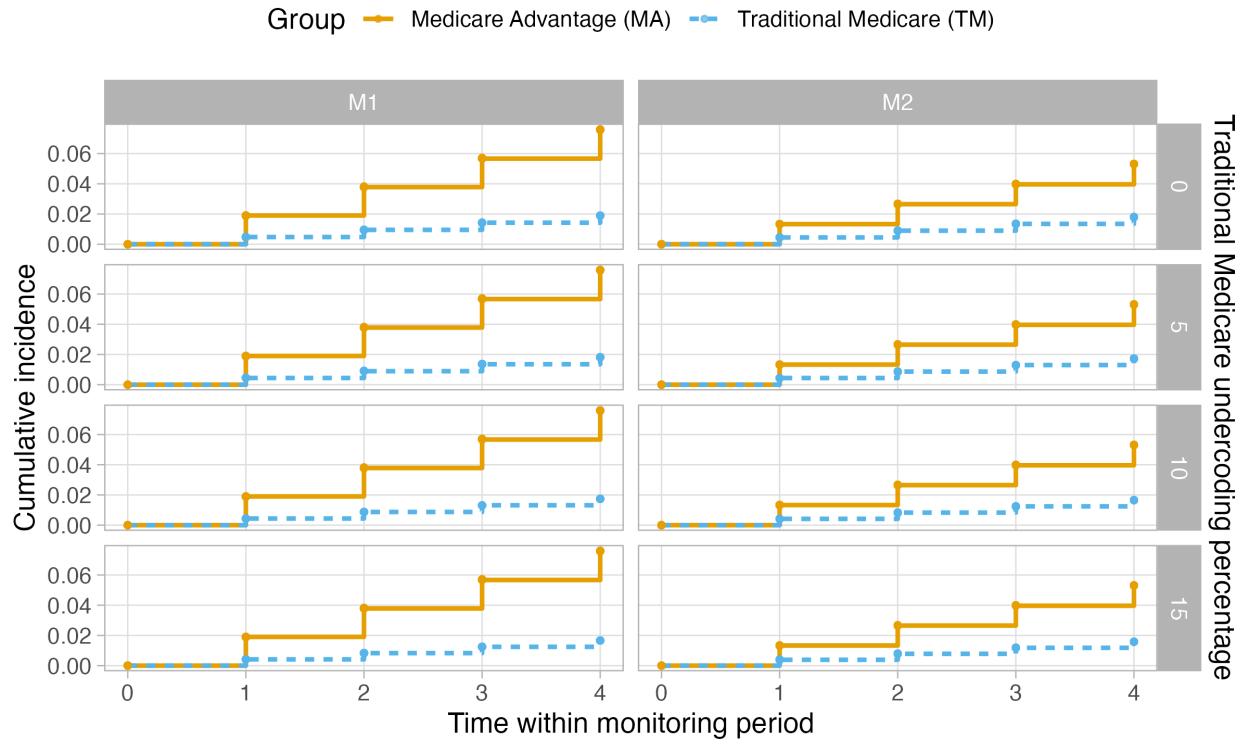


Figure 2: **Cumulative incidence functions for the Specified Heart Arrhythmias Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 30% any-available MA upcoding.** Specified heart arrhythmias corresponds to HCC238, which does not have any competing events. For this HCC, 30% of any-available individuals in the MA group are upcoded and 5% of any-available individuals in the TM comparison group are upcoded. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.



**Figure 3: Cumulative incidence functions for the ‘Dementia, Severe’ Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 20% lower severity MA upcoding.** ‘Dementia, Severe’ corresponds to HCC125, which has lower severity HCCs ‘Dementia, Moderate’ (HCC126) and ‘Dementia, Mild or Unspecified’ (HCC127). 20% of any individuals in the MA group who were previously coded with either HCC126 or HCC127 are upcoded and 5% of individuals in the TM comparison group are upcoded similarly. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

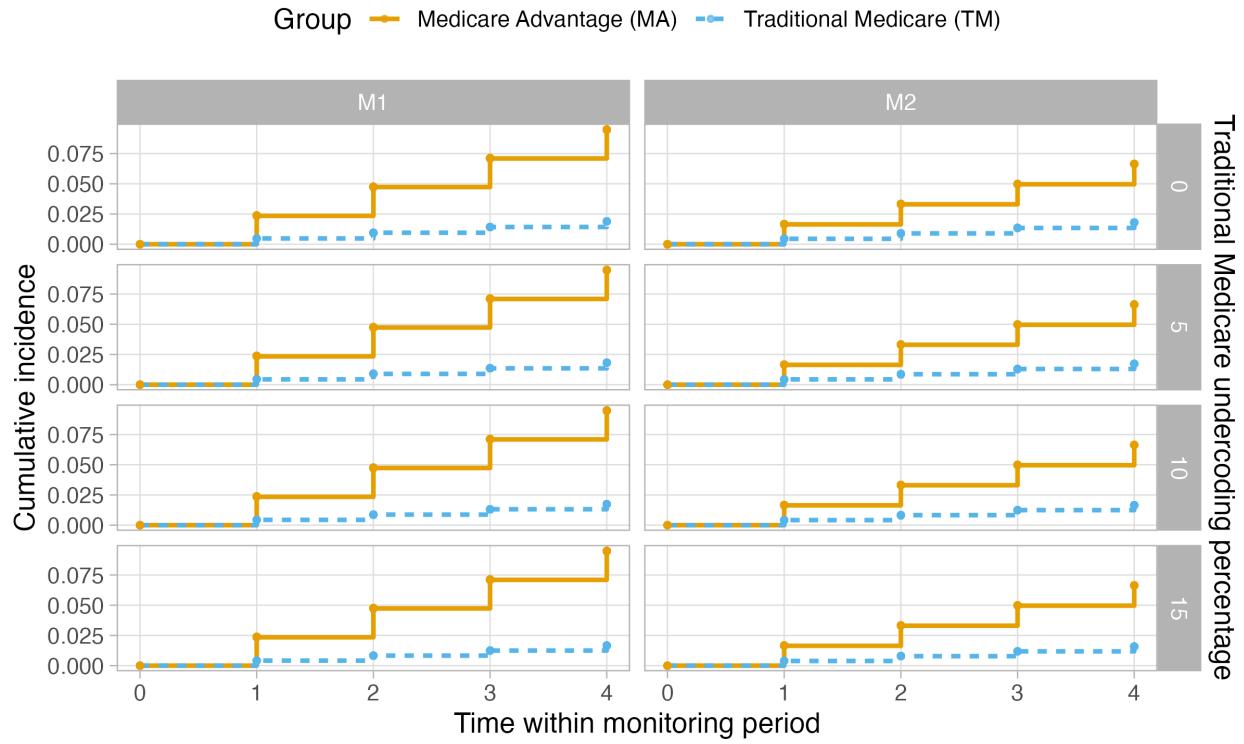
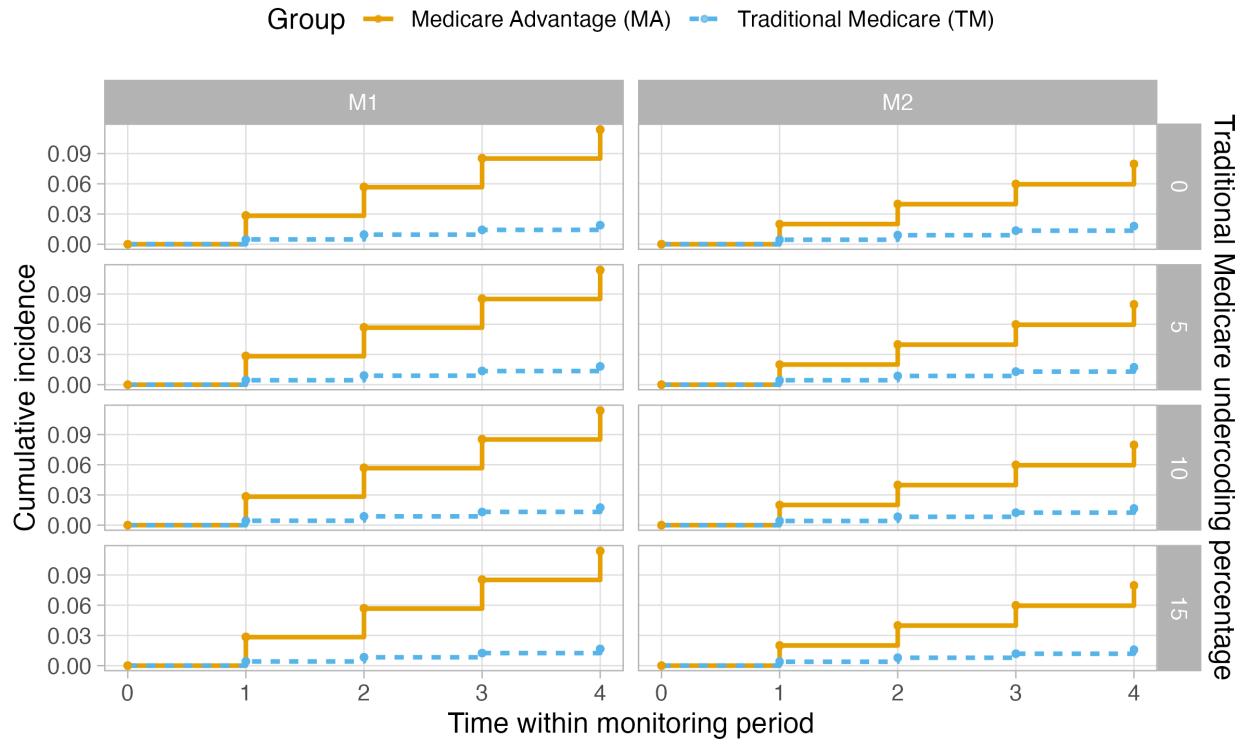


Figure 4: **Cumulative incidence functions for the ‘Dementia, Severe’ Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 25% lower severity MA upcoding.** ‘Dementia, Severe’ corresponds to HCC125, which has lower severity HCCs ‘Dementia, Moderate’ (HCC126) and ‘Dementia, Mild or Unspecified’ (HCC127). 25% of any individuals in the MA group who were previously coded with either HCC126 or HCC127 are upcoded and 5% of individuals in the TM comparison group are upcoded similarly. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.



**Figure 5: Cumulative incidence functions for the ‘Dementia, Severe’ Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups: 30% lower severity MA upcoding.** ‘Dementia, Severe’ corresponds to HCC125, which has lower severity HCCs ‘Dementia, Moderate’ (HCC126) and ‘Dementia, Mild or Unspecified’ (HCC127). 30% of any individuals in the MA group who were previously coded with either HCC126 or HCC127 are upcoded and 5% of individuals in the TM comparison group are upcoded similarly. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

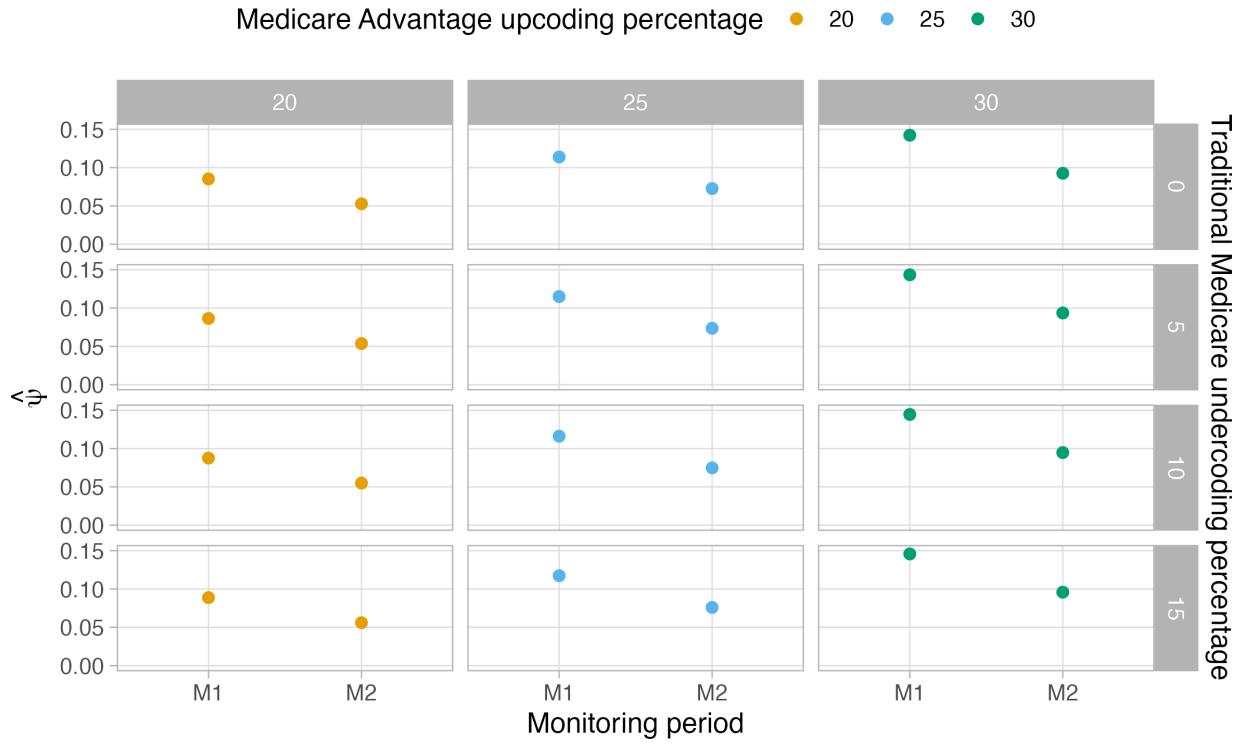


Figure 6: **Within-monitoring period  $\psi$  estimates for the ‘Dementia, Severe’ Hierarchical Condition Category (HCC) in simulated Medicare Advantage (MA) and Traditional Medicare (TM) groups.** ‘Dementia, Severe’ corresponds to HCC125, which has lower severity HCCs ‘Dementia, Moderate’ (HCC126) and ‘Dementia, Mild or Unspecified’ (HCC127). For this HCC, individuals in the MA group who were previously coded with either HCC126 or HCC127 are upcoded to varying degrees and individuals in the TM comparison group are upcoded 5% similarly. Given the large sample size, confidence intervals are very narrow and are therefore omitted as they cannot be distinguished visually.

### 3. Data availability statement

The simulated data for this study can be generated using code in the project repository: [https://github.com/StanfordHPDS/tte\\_estimation\\_medicare](https://github.com/StanfordHPDS/tte_estimation_medicare). Additional upcoding, undercoding, and baseline data can be simulated using the upcoding package (<https://github.com/StanfordHPDS/upcoding>). In line with program policies, non-summarized version 7 All of Us survey data used to derive baseline co-occurring HCCs can be accessed by authorized users via the All of Us Researcher Workbench only (<https://www.researchallofus.org/data-tools/workbench/>) and requires Registered Tier access. Code that was used for this project in the Researcher Workbench is in the above project repository.

### References

1. Centers for Medicare and Medicaid Services. *2024 MA Advance Notice.*; 2023. <https://www.cms.gov/files/document/2024-advance-notice-pdf.pdf>

2. Centers for Medicare and Medicaid Services. 2024 model software/ICD-10 mappings, 2024 midyear/final model software. Published online 2024. <https://www.cms.gov/medicare/health-plans/medicareadvtgsspecratestats/risk-adjustors/2024-model-software/icd-10-mappings>