A note on efficiency in estimation with monotonically missing at random data[*]

Jean-Louis Barnwell[†] and Saraswata Chaudhuri[‡]

August 10, 2023

Abstract

We study efficiency in estimation of parameters of generic target (sub) populations defined by the missingness pattern of monotonically missing at random data. Monotonic missingness is typically generated by attrition from multi-period studies if individuals never return after leaving. Then our target parameters describe functions of the counterfactuals that were unrealized due to attrition in different periods. We obtain the efficient influence functions for the target parameters, emphasizing on novel issues related to over identification, usability of sample units, and the information content of the missing at random conditions. A standard doubly robust estimator for the generic target parameter follows by equating to zero the sample analog of its efficient influence function that plugs in parametric or nonparametric estimators of the unknown nuisance parameters. It performs well and vastly outperforms other estimators in our simulation experiment and empirical illustration.

Keywords: Attrition; Efficiency; Monotonic MAR; Project STAR; Sub-populations.

[*] We are very grateful to the co-editor, two anonymous referees and Whitney Newey for their help in improving the paper. We also thank Francesco Amodio, Marine Carrasco, Daniel Farewell, Bryan Graham, Fabian Lange, Steven Lehrer, Thierry Magnac, Erica Moodie, Chris Muris, Tom Parker, Geert Ridder, Youngki Shin and various conference and seminar participants for helpful comments.

[†] Department of Economics, McGill University, Montreal, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

[‡] Corresponding author. Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

1 Introduction

Subjects/respondents often leave at various junctures of multi-period/phase studies/surveys. If they do not return then that creates a monotonically missing data set with respect to the original cohort of the study/survey. Monotonicity is reflected by the fact that the members of the original cohort that are observed in a later period are also observed in earlier periods. Equivalently, monotonicity is also reflected by the fact that the variables observed for those that left after an earlier period are also observed for those that left after later periods.

Attrition in the sample causes problems with statistical analysis. First, the sample's representativeness of the original population may be lost. Second, even if representativeness is restored by virtue of plausible assumptions such as missingness at random (selection on observables), the loss in data leads to imprecision in estimation; and, therefore, efficient estimation that optimally uses the remaining available information is of utmost importance.

Our paper is about efficiency in estimation with monotonically missing at random data. We build on the early work of Robins and Rotnitzky (1992), Robins et al. (1995), Rotnitzky and Robins (1995), Fitzgerald et al. (1996), Abowd et al. (2001), Wooldridge (2002), Nicoletti (2006), Wooldridge (2010), etc. in the biostatistics and econometrics literature extending them to sub-populations defined by the monotone pattern of missingness. Such sub-populations are interesting because they reflect the attrition behavior of economic agents; e.g., agents left school or job or marriage after period one, after period two, ..., never left.

To set the benchmark that any regular estimator should strive to reach, we obtain the efficiency bound for estimating parameters in general moment restrictions models. Our proposed estimator can reach this bound, and belongs in the class of two-step estimators satisfying double robustness with respect to the underlying nuisance parameters that can be estimated parametrically or nonparametrically. This class of estimators is well studied and known to be attractive in practice; see, e.g., Robins et al. (1994), Robins and Ritov (1997), Holcroft et al. (1997), Scharfstein et al. (1999), Bang and Robins (2005), Tsiatis (2006), Tan (2007), Cao et al. (2009), Rothe and Firpo (2019), Chernozhukov et al. (2018), etc.

Our results provide insights on the relation between the information content of the missing at random (MAR) assumption and the usability of the sample units toward estimation in sub-populations. The general (i.e., weakest) MAR assumption is that the hazard of leaving at any period does not depend on what would have happened afterwards once we condition on "all" that has already happened. Under this general MAR assumption, we show that if interest lies on those that left at the end of, e.g., period four, then those that left before period four are not usable for estimation. By contrast, we show that if it is plausible to strengthen this general MAR assumption by restricting the "all" in its conditioning set as in, e.g., Chaudhuri (2020), then more (not all) sample units for who the restricted "all" is observed and not just those that did not leave before period four become usable for estimation.

We also show that the efficiency bounds under the general MAR assumption coincide with that of particular augmented moment condition problems. A similar analysis in Chaudhuri (2020) was built on Graham (2011) that was based on an orthogonalization in Brown and Newey (1998). That can not work for sub-populations in our setup because the key nuisance parameters — the conditional hazards of leaving — are unknown. (Known/unknown did not matter for Graham (2011) since he focused on the full population; see Hahn (1998).) Here, on the other hand, we need to use the orthogonalization in Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2018), etc. for a unified treatment of full and sub populations.

This orthogonalization implies that, in theory, the asymptotic variance of an inverse probability weighted (IPW) estimator based on nonparametrically estimated nuisance parameters will equal the inverse of the efficiency bound under the general MAR condition. However, our simulations suggest that this theory could severely underestimate IPW's true variability (measured by Monte Carlo variance) even in very large samples when, unlike in Hirano et al. (2003), Chen et al. (2008), Graham (2011), etc., we move beyond the single level of missingness. Our simulations also suggest that even the more conservative (in finite samples) formula for asymptotic variance in the spirit of Ackerberg et al. (2012) can be a poor approximation underestimating IPW's true variability in small samples. Hence IPW is not our recommended estimator. On the other hand, at least in our simulations we do not see either of these two problems with our proposed estimator.

We also note that while this orthogonalization from Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2018), etc. provides valid influence functions, it may not lead to semiparametric efficiency in general. Its claim to efficiency is solely based on a given moment function (e.g. IPW) involving unknown nuisance parameters that are nonparametrically exactly identified by a second set of moments, and on no additional information like the MAR assumption. In our setup, however, semiparametric efficiency is tied to the strength of the MAR assumption. While the general MAR assumption turns out to not contain any relevant information in this context, we show that when we strengthen that assumption then the said orthogonalization cannot reach the resulting efficiency bound. This suggests that while such orthogonalizations are obviously very useful, it is still important to consider all the available information to obtain the semiparametric efficiency bound that follows from it.

Finally, an important feature of our paper is that we obtain the efficiency results for parameters defined by over identifying moment restrictions. This is not common in this literature; Chen et al. (2008) is among notable exceptions. To our understanding, the characterization of the tangent set in Chen et al. (2008) may be incomplete because over identification is not explicitly used for that.[1] We show that the efficiency results in Chen et al. (2008) still hold. We also show that the efficiency results in Chaudhuri (2020) under — (i) the general MAR with planned (known) conditional hazards or (ii) his convenient MAR — can be extended to over identifying moment restrictions. On the other hand, under our setup it seems that a complete characterization of the tangent set hinders a seamless transition of the efficiency results for certain (not all) sub-populations between just and over identification. We provide a detailed treatment of this issue as it seems to be less appreciated (at least we did not know before an anonymous referee for Chaudhuri (2020) pointed it out).

Our paper proceeds as follows. Section 2 lays out the theoretical framework guided by

[1] We are very grateful to an anonymous referee for Chaudhuri (2020), Patrik Guggenberger and Whitney Newey for their help with this issue. Any error is of course only our responsibility.

an empirical motivation based on the attrition behavior of students from the widely studied, attrition-infested Project STAR experiment. Section 3 presents the core theory – efficiency bound, efficient influence function, over identification, and the information content of the MAR assumption – by relating them with the literature. Section 4 presents the estimator and a sketch of its properties under parametric (mis)specification and nonparametric specification of the nuisance parameters. The asymptotic theory of such estimators is well studied and is certainly not our contribution; the sketch is presented only for completeness. Section 5 presents an elaborate empirical illustration of the benefits of the proposed estimator's precision in drawing substantive conclusions on the effect of small class size across dimensions induced by the attrition behavior of students from Project STAR. Section 6 concludes.

All the proofs are collected in Supplemental Appendix A. Complementing the theory in our paper, we present in Supplemental Appendix B a Monte Carlo experiment demonstrating excellent small-sample properties of our proposed estimator. The experiment also suggests that the promise of efficiency made by the theory for the competing IPW estimators based on nonparametric estimation of nuisance parameters may not realize even in very large samples.

2 Empirical motivation and the theoretical framework

2.1 Empirical motivation

Tennessee's Student/Teacher Achievement Ratio experiment, also known as Project STAR, has been extensively used to study the effect of small class size on future outcomes for the students; see, e.g., Hanushek (1999), Krueger (1999), Krueger and Whitmore (2001), Ding and Lehrer (2010), Chetty et al. (2011), etc. In Project STAR, students enrolling in grade K of 79 participating schools in the 1985-1986 school year were randomly assigned to three types of classes: small classes (13-17 students per teacher), regular classes (22-25 students per teacher), and regular classes with a full-time teacher's aide (22-25 students per teacher). The literature on Project STAR typically does not differentiate between the latter two class types, and we will follow that here and refer to them jointly as "not-small" classes.

We use the well-known and publicly available Project STAR data (Achilles et al. (2008)) containing characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students and their normalized reading and math scores from grade K to grade 3 or to a lower grade until which they stayed with a Project STAR school.[2]

Many students — 701 out of 1493 (47%) from small classes and 1725 out of 3477 (49.6%) from not-small classes — entering Project STAR schools in grade K did not stay until the project ended, i.e., until the end of grade 3.[3] See Table 1. For simplicity of illustration, we further exclude from our sample the very small percentage of students who switched classes.[4]

	Randomized to small class			Randomized to not-small class		
After	Stayed	Left STAR	Switched	Stayed	Left STAR	Switched
grade	in small	school	to not-small	in not-small	school	to small
K	1004	410	79 (5.3%)	2230	1047	200 (5.8%)
1	798	188	18 (1.8%)	1674	481	75 (3.4%)
2	672	103	23 (2.9%)	1392	197	85 (5.1%)

Table 1: Number of students in our sample by their switching class type or leaving Project STAR dynamics at the end of each grade conditional on staying until the end of that grade in their initially assigned class. The switcher % inside the parentheses are with respect to the class-type specific row total, e.g., $100 \times 79/(1004 + 410 + 79) \approx 5.3$

This attrition makes the scores of a student in a grade unobserved/counterfactual if the student left before completing the grade. Consequently, many of the grade-specific average scores that researchers compare to estimate the effect of small classes are unavailable. To fix ideas, consider the reading scores reported in Table 2. Note that the grade-specific average reading scores in small or non-small classes are the weighted average of the elements of

[2] We work with normalized scores for the sake of interpretation. For example, the normalized reading score is the demeaned and standardized reading score of each student at each grade based on that grade's mean and standard deviation of reading scores of students across all participating Project STAR schools.

[3] In the original data set, 917 out of 1900 (48.3%) from small classes and 2139 out of 4425 (48.3%) from not-small classes entering Project STAR schools in grade K did not stay until the end of grade 3. For simplicity of the illustration, we construct our working sample by dropping from this original data set students: (i) who did not enroll in Project STAR schools in grade K in 1985 but enrolled in grades 1, 2 and 3 in the next three years, or (ii) who left Project STAR schools after grades K or 1 or 2 but came back in the subsequent years during the experiment, or (iii) with incidental missing (relevant) variables when the missingness is unrelated to attrition, or (iv) with invalid test scores (see, e.g., p. 151 of Hanushek (1999)).

[4] Only 18 and 23 students switched from small class after grades 1 and 2 respectively. These numbers are too small for any analysis without extremely stringent restrictions on models for the switching behavior. We do not know enough to impose such stringent restrictions and hence exclude the switchers from our analysis.

that grade's column in Table 2 with weights proportional to the corresponding number of students, e.g., for grade K in small class it is $(-.19 \times 410 - .14 \times 188 - .09 \times 103 + .45 \times 672)/(410 + 188 + 103 + 672) \approx .14$. The grade-specific averages are unavailable (marked by "?") except in grade K because attrition starts after grade K.

Left STAR	Randomized to small class					Randomized to not-small class				
school at the	Number of	Reading score in grade				Number of	Reading score in grade			
end of grade	students	K	1	2	3	students	K	1	2	3
K	410	-.19	x	x	x	1047	-.36	x	x	x
1	188	-.14	-.19	x	x	481	-.27	-.54	x	x
2	103	-.09	-.14	-.23	x	197	-.00	-.22	-.31	x
3 (Never left)	672	.45	.50	.47	.44	1392	.20	.33	.30	.22
Average Score		.14	?	?	?		-.07	?	?	?

Table 2: Observed and unobserved normalized reading scores by attrition behavior of students from their initially assigned classes. If the full population is of interest then the number of levels of missingness in any grade's score is the number of x in that grade's column.

Naively imputing these grade-specific averages by the "Never left" category would be extremely misleading for both small (.14 by .45) and not-small (-.07 by .20) classes in grade K. (We could compare since grade K scores are actually observed for all.) Therefore, naive imputation based on the Never left category would possibly be misleading as well for grades 1, 2 and 3, where some sort of imputation is actually required. Interestingly, such imputations are less misleading when we compare the difference between the averages of grade K reading score in small and not-small classes: (.45 - .20) - (.14 - (-.07)) = .25 - .21 = .04 — the effect of attrition largely cancels out, which can be seen as a type of "common trend" phenomenon.[5]

However, the investigation cannot end here as there are two outstanding questions. First, will the same phenomenon emerge from the scores in grades 1, 2 and 3? Second, are any of those differences between small and not-small classes going to be statistically significant?

The first question is not answerable without assumptions on the mechanism of attrition because it involves comparing counterfactual means with the scores of the Never left category.

[5] Similar observations have been made repeatedly in economics; see, e.g., the Special Issue: "Attrition in Longitudinal Surveys" in the Journal of Human Resources (1998) where one observes that big distortions of group means due to attrition often vanish in the results of regression, i.e., for difference in group means.

We do not have anything original to say in this regard and will work under a very general MAR (selection on observables) assumption with a very flexible model specification for it.

On the other hand, our paper is about efficiency in estimation and is devised to address the second question. Under a general MAR assumption, we will estimate the counterfactual means with likely most precision and check if the concerned differences are statistically significant. (Section 5 will present strong evidence that such differences are significant.)

To efficiently estimate the grade-specific counterfactual mean we will need to efficiently estimate the attrition-category-specific counterfactual means in each grade, i.e., the ones that are marked by "x" in Table 2. These are examples of what we mean by sub-population-specific parameters where the sub-populations partition the full population by the attrition behavior of the students (population units). Such sub-population-specific parameters are obviously important for many other purposes including as descriptive statistics, and we will make use of them in various ways in the empirical illustration in Section 5.

2.2 Theoretical framework

Let $Z := (Z_1', \ldots, Z_R')'$ where Z_r is a $d_r \times 1$ random vector and $\sum_{r=1}^{R} d_r$ is finite. Let C be a random variable with support $\{1, \ldots, R\}$. Let $T_C(Z)$ be a transformation defined as $T_r(Z) := (Z_1', \ldots, Z_r')'$ for $r = 1, \ldots, R$. The notation is standard; see, e.g., Tsiatis (2006). Z_j's may have common elements, e.g., time invariant variables, and empirical practice (coding, etc.) should ensure that they are not counted in the $T_r(Z)$'s more than once.

Let $O := (C, T_C'(Z))'$ denote what is observed for a unit in the sample.

Consider the Project STAR example. This is an $R = 4$ period study where grade K is period 1,..., and grade 3 is period 4. Z_r are the variables — characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students and their normalized reading and math scores in period r — that are observed in period r. $T_r(Z)$ is the cumulative history of the Z_r variables (some of which may be time-invariant) observed until and including period r. If a unit (student) leaves after period $j \in \{1, \ldots, R\}$, then its $C = j$ and we only observe $T_j(Z)$ for it. $C = R$ is the same as never leaving (some denote this as $C = \infty$).

We maintain the general MAR (selection on observables) assumption that:

$$P(C = r | T_R(Z), C \geq r) = P(C = r | T_r(Z), C \geq r) \text{ for } r = 1, \ldots, R-1. \tag{1}$$

Since $T_r(Z)$ is observable when $C \geq r$, (1) imposes that the conditional hazard $P(C = r | T_R(Z), C \geq r)$ at period r does not depend on the unobservables $Z_{r+1}, \ldots, Z_R$ once conditioned on the observables $T_r(Z)$. (1) is the MAR assumption in the sense of Rubin (1976).

Plausibility of MAR depends on the context. MAR has been widely used in studies on attrition especially if, as in our paper, the missingness is monotone.[6] Ding and Lehrer (2010) (and, less explicitly, Krueger (1999)) assumed MAR for attrition in the Project STAR data.

Generalizing the nomenclature introduced in Section 2.1, we refer to the underlying population of $O := (C, T'_C(Z))'$ as the full population. We refer to the partition of this full population by the values taken by C as sub-populations; e.g., sub-population r is the underlying population from which units with $C = r$ can be viewed as randomly drawn. There are R unitary sub-populations indexed by $r = 1, \ldots, R$. Unions of unitary sub-populations form a composite sub-population, e.g., $C \in \{1, 2\}$, or the full population $C \in \{1, \ldots, R\}$.

Under the general MAR condition in (1), the unconditional distribution of Z may not be the same as the distribution of Z conditional on $C = r$ for $r = 1, \ldots, R$, i.e., the sub-populations are possibly heterogeneous. In the example from Section 2.1 where the sub-populations are defined by the attrition categories based on the timing of attrition, this means that the distribution of the (potential) grade 3 reading scores may not be same for those who left after grade K and those who left after grade 2 and those who never left.

We will work with a generic target sub-population $C \in \{a, \ldots, b\}$, denoted for brevity by $a \leq C \leq b$ or $[a, b]$, for $a \leq b$ and $a, b \in \{1, \ldots, R\}$. If $a = b = r$ then this is the underlying

[6] If the missingness is non-monotone, then MAR or selection on observables is unrealistic since the choice to return could depend on unobservables, i.e., on what happened when the individual was out of the study; see, e.g., Gill and Robins (1997), Gill et al. (1997), Robins and Gill (1997) and Vansteelandt et al. (2007). That would be a case of selection on unobservables. Hoonhout and Ridder (2019) compare various selection on unobservables conditions with MAR in a multi-period context. We do not contribute to that literature.

unitary sub-population from which the units who left at the end of period r can be viewed as randomly drawn. If $a < b$ then this is the composite sub-population for the units who left in the periods $a, a+1, \ldots, b$. If $a = 1$ and $b = R$ then this is the full-population.

Denote the distribution of Z in the target population by $F_{Z|(a \leq C \leq b)}(z)$. This is the weighted average of the distributions of Z in sub-populations $a, \ldots, b$ with weights $P(C = j)/P(a \leq C \leq b)$ for $j = a, \ldots, b$. We will define the parameter of interest as a finite dimensional feature of $F_{Z|(a \leq C \leq b)}(z)$. Accordingly, consider a function $m(Z; \beta) : \text{Support}(Z) \times \boldsymbol{\mathcal{B}} \mapsto \mathbb{R}^{d_m}$, $\beta \in \boldsymbol{\mathcal{B}} \subset \mathbb{R}^{d_\beta}$ and $d_\beta \leq d_m$. Then, for a given $a, b \in \{1, \ldots, R\}$ with $a \leq b$, define the parameter value of interest $\beta^0_{[a,b]}$ by an over identifying system of moment restrictions as:

$$E[m(Z; \beta) \mid a \leq C \leq b] = 0 \text{ for } \beta \in \boldsymbol{\mathcal{B}} \;\text{ if and only if }\; \beta = \beta^0_{[a,b]}. \tag{2}$$

$m(Z; \beta)$ can depend on any element of Z; e.g., reading score in grade K or 1 or 2 or 3. If the least frequently observed element of Z that is involved in $m(Z; \beta)$ belongs in Z_k for some $k = 1, \ldots, R$ then exactly the same analysis in the sequel will still apply but with a different observability indicator $\bar{C}$ instead of C where $\bar{C} := k$ if $C \geq k$ and $\bar{C} := C$ otherwise.

We will also maintain the following assumptions that are standard in this literature.

Assumption A:

(A1) The observed sample units $\{O_i := (C_i, T'_{C_i}(Z_i))'\}_{i=1}^n$ are i.i.d. copies of $O := (C, T'_C(Z))'$.

(A2) $P(C = R | T_R(Z))$ is bounded away from zero almost surely $T_R(Z)$.

(A3) $M_{[a,b]}$ is a $d_m \times d_\beta$ finite matrix of full column rank where $M_{[a,b]} := M_{[a,b]}(\beta^0_{[a,b]})$ and $M_{[a,b]}(\bar{\beta}) := \left\{ \frac{\partial}{\partial \beta'} E\left[m(Z; \beta) | \, a \leq C \leq b \right] \right\}_{\beta = \bar{\beta}}$ at any $\bar{\beta} \in \boldsymbol{\mathcal{B}}$ where it exists.

Remark: (A1) rules out dependence and heterogeneity across sample units when viewed as random draws from O. (A2) imposes the bounded away from zero condition instead of only $P(C = R | T_R(Z)) > 0$ to avoid the "limited overlap" problem; see, e.g., Khan and Tamer (2010). (A3) gives local identification of $\beta^0_{[a,b]}$; it allows for non-smooth $m(Z; \beta)$ but requires the expectation to be differentiable with respect to β; see, e.g., Chen et al. (2008).

3 The efficiency results

3.1 Efficiency bound and efficient influence function

Writing $T_r(Z)$ as T_r for $r = 1, \dots, R$, let us first introduce the key quantities for this section. Define:

$$\varphi_{[a,b]}(O;\beta) := \sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \varphi_{[j,j]}(O;\beta) \tag{3}$$

for the sub-population $[a,b]$ as the weighted average of the unitary sub-population quantities:

$$\varphi_{[j,j]}(O;\beta) := \sum_{r=j+1}^{R} I(C \ge r)\omega_{r,j}(T_{r-1})\left(E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]\right) + \frac{I(C=j)}{P(C=j)} E[m(Z;\beta)|T_j] \tag{4}$$

that are feasible for each $\beta \in \mathcal{B}$ based on the observed data because of the equality in (5):

$$\omega_{r,j}(T_{r-1}) := \frac{P(C=j|T_j)}{P(C=j)P(C \ge r|T_{r-1})} = \frac{P(C=j|T_j, C \ge j)}{P(C=j)\prod_{k=j}^{r-1}\left[1 - P(C=k|T_k, C \ge k)\right]}. \tag{5}$$

Under regularity conditions, the weighted average representation of $\varphi_{[a,b]}(O;\beta)$ implies:

$$\begin{aligned}
\frac{\partial}{\partial \beta'} E\left[\varphi_{[a,b]}(O;\beta)\right] &= \sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \frac{\partial}{\partial \beta'} E\left[\varphi_{[j,j]}(O;\beta)\right] = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)|\, C=j\right] \\
&= \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)|\, a \le C \le b\right], \quad \text{and} \\
Var\left(\varphi_{[a,b]}(O;\beta)\right) &= \sum_{j=a}^{b}\sum_{k=a}^{b} \frac{P(C=j)P(C=k)}{P^2(a \le c \le b)} Cov\left(\varphi_{[j,j]}(O;\beta), \varphi_{[k,k]}(O;\beta)\right).
\end{aligned}$$

The covariance ($j \neq k$) terms in the composite (sub-)populations simplify when $a = 1, b = R$.

Lemma 1 *In the case of the full population $a = 1, b = R$ the above representation gives:*

$$\begin{aligned}
\varphi_{[1,R]}(O;\beta) &= \sum_{r=2}^{R} \frac{I(C \ge r)}{P(C \ge r|T_{r-1})} \left(E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]\right) + E[m(Z;\beta)|T_1], \\
Var\left(\varphi_{[1,R]}(O;\beta)\right) &= \sum_{r=2}^{R} E\left[\frac{Var\left(E[m(Z;\beta)|T_r]|T_{r-1}\right)}{P(C \ge r|T_{r-1})}\right] + Var\left(E[m(Z;\beta)|T_1]\right).
\end{aligned}$$

Equipped with these key quantities, we will now present the main result of our paper.

Proposition 2 *Let the MAR condition in* (1)*, the moment restrictions in* (2) *and assumption A hold. Let* $V_{[a,b]} := Var(\varphi_{[a,b]}(O;\beta^0_{[a,b]}))$ *be a finite and positive definite matrix.*[7] *Then the semiparametric efficiency bound for* $\beta^0_{[a,b]}$ *is given by* $\Omega_{[a,b]} := M'_{[a,b]}V^{-1}_{[a,b]}M_{[a,b]}$*:*
(i) when $a=1, b=R$ *(full population) or* $a=b$ *(unitary sub-populations);*
(ii) when $a,b\in\{1,\dots,R\}$ *with* $a\le b$*, if additionally* $\beta^0_{[a,b]}$ *is just-identified, i.e.,* $d_m=d_\beta$*.*
A regular estimator $\widehat{\beta}_{[a,b]}$ *whose asymptotic variance equals* $\Omega^{-1}_{[a,b]}$ *has the asymptotically linear representation (with obvious cancellations giving* $\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]} = M^{-1}_{[a,b]}$ *when* $d_m=d_\beta$*):*

$$\sqrt{n}(\widehat{\beta}_{[a,b]}-\beta^0_{[a,b]}) = -\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta^0_{[a,b]})+o_p(1).$$

Remarks: First, Proposition 2 covers the well-known special cases found in the literature. $R=2$ with $a=b=1$ or $a=1,b=2$ covers Theorem 1 of Chen et al. (2008); also see Robins et al. (1994). $a=1,b=R>2$ gives the full-population result like Robins and Rotnitzky (1992), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997).

Second, few papers in this literature allow for $d_m>d_\beta$, i.e., over identifying restrictions for $\beta^0_{[a,b]}$. Chen et al. (2008) is among notable exceptions. However, it is possible that the characterization of the tangent set there (and similar papers) may be incomplete because over identification is not explicitly used for that. Proposition 2(i) shows that Chen et al. (2008)'s results (Chen et al. (2008) worked with $R=2$ with $a=b=1$ or $a=1,b=2$) still hold. Additionally, in Section 3.2 we also extend the main efficiency results in Chaudhuri (2020) (also a generalization of Chen et al. (2008)) to the case of over identifying restrictions.

[7] While it is easier to think of primitive conditions for positive definiteness of $Var\left(\varphi_{[a,b]}(O;\beta)\right)$ when $a=b$ or $a=1,b=R$, we maintain positive definiteness of $Var\left(\varphi_{[a,b]}(O;\beta^0_{[a,b]})\right)$ generally. Writing $\varphi_{[s,t]}(O;\beta)$ as $\varphi_{[s,t]}$ for $s,t=1,\dots,R$ and $m(Z;\beta)$ as m for brevity, the components of $Var\left(\varphi_{[a,b]}\right)$ can be expressed as follows. For $j=a,\dots,b$ and $k=a,\dots,j-1$: $Var\left(\varphi_{[j,j]}\right)=\sum_{r=j+1}^{R}E\left[\Delta_{r,j}|C=j\right]+Var\left(\frac{I(C=j)}{P(C=j)}E[m|T_j]\right)$ and $Cov\left(\varphi_{[j,j]},\varphi_{[k,k]}\right)=E\left[\sum_{r=j+1}^{R}\Delta_{r,j}+\sum_{r=k+1}^{j}\nabla_{r,j,k}\middle|C=j\right]+Cov\left(\frac{I(C=j)}{P(C=j)}E[m|T_j],\frac{I(C=k)}{P(C=k)}E[m|T_k]\right)$ where, again for simplicity, we have used the notation $\Delta_{r,j}:=\omega_{r,j}(T_{r-1})Var\left(E[m|T_r]|T_{r-1}\right)$ for $r=j+1,\dots,R$, and $\nabla_{r,j,k}:=\omega_{r,k}(T_{r-1})E[m|T_j]\left(E[m|T_r]-E[m|T_{r-1}]\right)'$ for $r=k+1,\dots,j$. If, e.g., $a=b=j$ then primitive conditions for the positive definiteness of $Var\left(\varphi_{[j,j]}\right)$ can be guided by its expression above.

Third, over identification is not innocuous in our general framework. Under just identification, the efficiency bound in Proposition 2 applies to any $a, b \in \{1, \dots, R\}$ with $a \leq b$. However, in case of over identification the efficiency bound result is for the full population ($a = 1, b = R$) and all R unitary sub-populations ($a = b$) but not for generic composite sub-populations $[a, b]$'s. Unlike in Chaudhuri (2020), here the over identifying restrictions for $\beta^0_{[a,b]}$ impose restrictions on the tangent set that do not seem to be satisfied for generic $[a, b]$'s by the influence function presented in the proposition. Since it seems less appreciated, we utilize Section 3.2 to be explicit about the restrictions imposed by over identification.

Fourth, the weighted average representation of $\varphi_{[a,b]}(O_i; \beta^0_{[a,b]})$ in (3), that follows from the representation $E[m(Z;\beta)|a \leq C \leq b] = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} E[m(Z;\beta)|C = j]$, presents an easy way of combining the efficient estimators for the unitary sub-populations to obtain the efficient estimator for the composite sub-population $[a, b]$ under just identification $d_m = d_\beta$:

$$\sqrt{n}\left(\widehat{\beta}_{[a,b]} - \sum_{j=a}^{b}\left[M^{-1}_{[a,b]}\frac{P(C=j)}{P(a \leq C \leq b)}M_{[j,j]}(\beta^0_{[a,b]})\right]\widehat{\beta}_{[j,j]}\right) = o_p(1)$$

where the weights for the $\widehat{\beta}_{[j,j]}$'s add up to the identity matrix since $M_{[a,b]} = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} M_{[j,j]}(\beta^0_{[a,b]})$.[8] Dardanoni et al. (2011), Abrevaya and Donald (2017), Muris (2020) and others also considered combining estimators or moment restrictions in similar contexts with missing data.

Fifth, each $\varphi_{[j,j]}(O; \beta^0_{[a,b]})$ is doubly robust to the misspecification of the two sets of unknown nuisance parameters: the conditional hazards $P(C = r|T_r, C \geq r)$ and the conditional expectations $E[m(Z;\beta)|T_r]$ for the various r's. Therefore, the representation of $\varphi_{[a,b]}(O; \beta^0_{[a,b]})$ in (3) implies that $\varphi_{[a,b]}(O; \beta^0_{[a,b]})$ also satisfies such double robustness. $\varphi_{[j,j]}(O; \beta^0_{[a,b]})$ is robust

[8] To see this result write the weights $M^{-1}_{[a,b]}\frac{P(C=j)}{P(a \leq C \leq b)}M_{[j,j]}(\beta^0_{[a,b]})$ as A_j for brevity and note that:

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{[a,b]} - \beta^0_{[a,b]}) &= -M^{-1}_{[a,b]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta^0_{[a,b]}) + o_p(1) = -M^{-1}_{[a,b]}\sum_{j=a}^{b}\frac{P(C=j)}{P(a \leq C \leq b)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[j,j]}(O_i;\beta^0_{[a,b]}) + o_p(1)\\ &= \sum_{j=a}^{b}A_j\left[-M^{-1}_{[j,j]}(\beta^0_{[a,b]})\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[j,j]}(O_i;\beta^0_{[a,b]}) + o_p(1) = \sum_{j=a}^{b}A_j\sqrt{n}\left(\widehat{\beta}_{[j,j]} - \beta^0_{[a,b]}\right) + o_p(1).\end{aligned}$$

The result follows since $\beta^0_{[a,b]}$ on both sides cancels out as $\sum_{j=a}^{b} A_j = I_{d_m} = I_{d_\beta}$ implies $\beta^0_{[a,b]} = \sum_{j=a}^{b} A_j\beta^0_{[a,b]}$.

to the misspecification of the $P(C = r|T_r, C \geq r)$'s under (1) since if we take expectation after replacing each $P(C = r|T_r, C \geq r)$ in (4) (precisely, (5)) by arbitrary scalar functions of T_r, we still obtain $E[m(Z; \beta)|C = j]$ if the expectation exists. To see that $\varphi_{[j,j]}(O; \beta^0_{[a,b]})$ is also robust to the misspecification of the $E[m(Z; \beta)|T_r]$'s, rewrite (4) as:

$$\begin{aligned}
\varphi_{[j,j]}(O;\beta) := \; & I(C = R)\omega_{R,j}(T_{R-1})m(Z;\beta) \\
& + \sum_{r=j+1}^{R-1} \left[\frac{I(C \geq r)}{P(C \geq r|T_{r-1})} - \frac{I(C \geq r+1)}{P(C \geq r+1|T_r)} \right] \frac{P(C = j|T_j)}{P(C = j)} E[m(Z;\beta)|T_r] \\
& + \left[\frac{I(C = j)}{P(C = j)} - \frac{I(C \geq j+1)}{P(C \geq j+1|T_j)} \frac{P(C = j|T_j)}{P(C = j)} \right] E[m(Z;\beta)|T_j],
\end{aligned} \tag{6}$$

replace each $E[m(Z; \beta)|T_r]$ in (6) by arbitrary d_m dimensional functions of T_r, take expectation while noting that $P(C \geq r|T_r) = P(C \geq r|T_{r-1})$ (see Lemma 9), and finally see that (1) gives the expectation as $E[I(C = R)\omega_{R,j}(T_{R-1})m(Z; \beta)] = E[m(Z; \beta)|C = j]$ (see Lemma 5) if the expectation exists. This is double robustness with respect to misspecification of nuisance parameters; see Robins et al. (1994), Robins and Ritov (1997), Scharfstein et al. (1999), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2019), Chernozhukov et al. (2018) etc. We do not contribute to the double robustness literature but rather use it to motivate the estimating function for $\beta^0_{[a,b]}$ in Section 4 based on $\varphi_{[a,b]}(O; \beta^0_{[a,b]})$.

Sixth, the expression for the $\varphi_{[j,j]}(O; \beta)$'s in (4) or (6) tells us that if $a \geq 2$ then the units with $C < a$ do not contribute to the estimation of the target $\beta^0_{[a,b]}$.[9] We note that this is an artifact of the general MAR condition in (1). Units with $C < a$ can contribute to the efficient estimation of $\beta^0_{[a,b]}$ if it is plausible to strengthen the MAR condition. A concrete example can be found in Proposition 4 below adopted for extension from Chaudhuri (2020). (This example of strengthened MAR is revisited in Section 3.3 to caution against sub-optimal use of sample units in case of over identification of the nuisance conditional hazards.) In extreme contrast, Proposition 3 below adopted for extension from Chaudhuri (2020) shows that all sample units are usable for all target $\beta^0_{[a,b]}$'s if the conditional hazards are actually known.

[9] Thus, generalizing the caption of Table 2, if Z_k is the least frequently observed element of Z that is involved in $m(Z; \beta)$ then the effective level of missingness is $\max\{0, k - a\}$ under the MAR condition in (1).

3.2 Over identification of $\beta^0_{[a,b]}$: restriction on the tangent set

Let f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. Their conditional counterparts are denoted similarly. Let $L^2_0(F)$ denote the space of mean-zero, square integrable functions with respect to F.

We will first characterize the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data $O = (C', T'_C(Z))'$. (We will then impose on it the restrictions due to over identification.) Consider a regular parametric sub-model indexed by a parameter η for the distribution of O. The log of this distribution is:

$$\log f_\eta(O) = \log f_\eta(Z_1) + \sum_{r=2}^{R} I(C \geq r) \log f_\eta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \log P_\eta(C = r|Z_1, \ldots, Z_r)$$

in terms of $(C, Z')'$. The score function with respect to η is, in terms of $(C, Z')'$,

$$S_\eta(O) = s_\eta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\eta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\eta(C = r|Z_1, \ldots, Z_r)}{P_\eta(C = r|Z_1, \ldots, Z_r)}$$

where $s_\eta(Z_1) := \frac{\partial}{\partial\eta} \log f_\eta(Z_1)$, $s_\eta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\eta} \log f_\eta(Z_r|Z_1, \ldots, Z_{r-1})$ for $r = 2, \ldots, R$, and $\dot{P}_\eta(C = r|Z_1, \ldots, Z_r) := \frac{\partial}{\partial\eta} P_\eta(C = r|Z_1, \ldots, Z_r)$ for $r = 1, \ldots, R$. The tangent set $\mathcal{T}$ is the mean square closure of all d_β dimensional linear combinations of $S_\eta(O)$ (see pp. 105-106, Newey (1990)) and can be expressed as:

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r)\nu_r(Z_1, \ldots, Z_r) + \sum_{r=1}^{R} I(C = r)\omega_r(Z_1, \ldots, Z_r) \tag{7}$$

where $\nu_1(Z_1) \in L^2_0(F(Z_1))$ and $\nu_r(Z_1, \ldots, Z_r) \in L^2_0(F(Z_r|Z_1, \ldots, Z_{r-1}))$ for $r = 2, \ldots, R$, and $\omega_r(Z_1, \ldots, Z_r)$ is any square integrable function of $Z_1, \ldots, Z_r$ for $r = 1, \ldots, R$.

This is typically how the tangent set is characterized in this literature (e.g., Chen et al. (2008)), but it does not take into account the additional restrictions imposed by over identification of $\beta^0_{[a,b]}$. Apart from the incompleteness in the proofs due to such omissions, it does seem that the additional restrictions will matter in our general setup with generic sub-populations $[a,b]$. Hence we will provide the details behind these additional restrictions.

For simplicity we will drop the subscript η from all quantities (e.g., in (8) below) evaluated at η^0 where η^0 is the "true" sub-model η, i.e., $f_{\eta^0}(O)$ is the actual distribution of the observed data. Note that, the moment restrictions in (2) give the following identity in η for given a,b:

$$0 \equiv E_\eta[m(Z;\beta^0_{[a,b]})|a \le C \le b] \equiv E_\eta\left[\frac{P_\eta(a \le C \le b|Z)}{P_\eta(a \le C \le b)} m(Z;\beta^0_{[a,b]})\right].$$

Differentiate it with respect to η under the integral at $\eta = \eta^0$, and use (1) and (2) to get:

$$0 = M_{[a,b]} \frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} + E\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z) + \frac{\dot{P}(a \le C \le b|T_b)}{P(a \le C \le b|T_b)} \right\}' \middle| a \le C \le b \right] \tag{8}$$

where $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$ (with abuse, we briefly revert to the T_r notation for brevity). Now, we note that (2) also gives the following identity in η for given a,b:

$$0 \equiv AE_\eta[m(Z;\beta^0_{[a,b]})|a \le C \le b] \equiv AE_\eta\left[\frac{P_\eta(a \le C \le b|Z)}{P_\eta(a \le C \le b)} m(Z;\beta^0_{[a,b]})\right]$$

for any A that is a full row rank $d_\beta \times d_m$ matrix such that $AM_{[a,b]}$ is nonsingular. Such an A always exists under our assumptions; e.g., $A = M'_{[a,b]} V^{-1}_{[a,b]}$. Therefore, following the same steps as in (8) and then solving for $\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'}$, we obtain that:

$$\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} = -\left(AM_{[a,b]}\right)^{-1} AE\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z) + \frac{\dot{P}(a \le C \le b|T_b)}{P(a \le C \le b|T_b)} \right\}' \middle| a \le C \le b \right],$$

which when substituted for in (8) gives (noting that $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$):

$$0 = \left(I_{d_\beta} - M_{[a,b]}\left(AM_{[a,b]}\right)^{-1} A\right) E\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1}) + \frac{\dot{P}(a \le C \le b|T_b)}{P(a \le C \le b|T_b)} \right\}' \middle| a \le C \le b \right] \tag{9}$$

Note that, (9) is trivially true under just identification $d_m = d_\beta$ since then $M_{[a,b]}\left(AM_{[a,b]}\right)^{-1} A = I_{d_\beta}$ by the definition of inverse. However, under over identification, (9) imposes restrictions on the quantities inside the expectations that must be reflected by the tangent set. Therefore, a complete characterization of the tangent set $\mathcal{T}$ in the case of over identification would augment what is defined in (7) such that its components additionally satisfy (10) if $[a,b]=[1,R]$ and satisfy (11) if $[a,b]\neq[1,R]$. Letting $B_{[a,b]} := \left(I_{d_\beta} - M_{[a,b]}\left(AM_{[a,b]}\right)^{-1}A\right)$,

- if the moment restrictions in (2) hold for $[a,b]=[1,R]$ then:

$$0 = B_{[1,R]}E\left[m(Z;\beta^0_{[1,R]})\sum_{r=1}^{R}\nu_r(Z_1,\dots,Z_r)'\right] \tag{10}$$

  as $\dot{P}_\eta(1\le C\le R|Z)=0$ in (9) since obviously $P_\eta(1\le C\le R|Z)\equiv 1$ for all η;[10]

- if the moment restrictions in (2) hold for $[a,b]\neq[1,R]$ then:

$$0 = B_{[a,b]}E\left[m(Z;\beta^0_{[a,b]})\left\{\sum_{r=1}^{R}\nu_r(Z_1,\dots,Z_r)+\sum_{r=a}^{b}\frac{P(C=r|Z_1,\dots,Z_r)}{P(a\le C\le b|Z_1,\dots,Z_b)}\omega_r(Z_1,\dots,Z_r)\right\}'\middle| a\le C\le b\right]. \tag{11}$$

Hence, $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ has to satisfy the restriction (10) or (11), as appropriate, to belong in $\mathcal{T}$ that is necessary for it to be the efficient influence function. Generalizing the literature, Proposition 2(i) showed that $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ satisfies the restriction when focus lies on the full population, i.e., $[a,b]=[1,R]$, or on the unitary sub-populations, i.e., $a=b$. Curiously, however, $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ does not seem to satisfy the restriction when $a\neq b$ but $[a,b]\neq[1,R]$, i.e., for composite sub-populations that are not the full population, and hence it is not efficient in that case although it remains a valid influence function since it satisfies the so-called "pathwise derivative" condition.

For completeness we note that a similar characterization of the tangent set allows us to extend the main efficiency results in Chaudhuri (2020) to the case of over identification. Those results work with strengthened MAR conditions but apply to remarkably more general

[10] We have not imposed enough structure on the $\omega_r(Z_1,\dots,Z_r)$'s to write (10) as a special case of (11). Other than here — restriction on tangent set due to over identification (that to our knowledge has not been covered in the MAR literature) — we presented full and sub-population analysis under the same framework instead of treating them separately as in; e.g., Hahn (1998), Hirano et al. (2003), Chen et al. (2008), etc.

target (sub-) populations λ. Precisely, Propositions 3 and 4 will concern a β^0_λ defined by the following moment restrictions: For any λ that is a subset of $\{1, \ldots, R\}$ including the full set, let

$$E[m(Z;\beta) \mid C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \;\text{ if and only if }\; \beta = \beta^0_\lambda. \tag{12}$$

Proposition 3 *Let the MAR condition in (1) and the moment restrictions in (12) hold. Let assumption A hold with $M_{[a,b]}$ in A3 replaced by $M_\lambda := E[\partial m(Z;\beta^0_\lambda)/\partial\beta' \mid C \in \lambda]$. Let $\bar{V}_\lambda := Var(\bar{\varphi}_\lambda(O;\beta^0_{[a,b]}))$ be a finite and positive definite matrix where:*

$$\bar{\varphi}_\lambda(O;\beta^0_\lambda) := \bar{\varphi}_{1,\lambda} + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_{r-1})}\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right) \quad \textit{with} \;\; \bar{\varphi}_{r,\lambda} := E\left[\left.\frac{P(C \in \lambda|T_r)}{P(C \in \lambda)} m(Z;\beta^0_\lambda)\right| T_r\right]$$

for $r = 1, \ldots, R$. If we additionally assume that $P(C = r|T_r, C \geq r)$ is known for $r = 1, \ldots, R-1$, i.e., the incompleteness is planned, then the semiparametric efficiency bound for β^0_λ is given by $\bar{\Omega}_\lambda := M'_\lambda \bar{V}^{-1}_\lambda M_\lambda$ and the efficient influence function is $-\bar{\Omega}^{-1}_\lambda M'_\lambda \bar{V}^{-1}_\lambda \bar{\varphi}_\lambda(O;\beta^0_{[a,b]})$.

The planned monotonic incompleteness condition was motivated in Chaudhuri (2020) as a cost cutting measure in survey designs. Another condition considered in Chaudhuri (2020) is a strengthened version of MAR, referred to as convenient MAR (CMAR), whereby:

$$P(C = r|Z, C \geq r) = P(C = r|T_1, C \geq r) \;\text{ for } r = 1, \ldots, R. \tag{13}$$

Proposition 4 *Let the moment restrictions in (12) and the CMAR condition in (13) hold. Let assumption A hold with $M_{[a,b]}$ in A3 replaced by $M_\lambda := E[\partial m(Z;\beta^0_\lambda)/\partial\beta' \mid C \in \lambda]$. Let $V^{CMAR}_\lambda := Var(\varphi^{CMAR}_\lambda(O;\beta^0_\lambda))$ be a finite and positive definite matrix where:*

$$\varphi^{CMAR}_\lambda(O;\beta^0_\lambda) := \frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta^0_\lambda)|T_1] + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_1)} \frac{P(C \in \lambda|T_1)}{P(C \in \lambda)} \left(E[m(Z;\beta^0_\lambda)|T_r] - E[m(Z;\beta^0_\lambda)|T_{r-1}]\right).$$

Then the semiparametric efficiency bound for β^0_λ is given by $\Omega^{CMAR}_\lambda := M'_\lambda [V^{CMAR}_\lambda]^{-1} M_\lambda$ and the efficient influence function is $-[\Omega^{CMAR}_\lambda]^{-1} M'_\lambda [V^{CMAR}_\lambda]^{-1} \varphi^{CMAR}_\lambda(O;\beta^0_{[a,b]})$.

3.3 IPW, variance adjustment, and information content of MAR:

Returning to the general MAR condition in (1), it is clear from (6) that $\varphi_{[j,j]}(O;\beta)$ is an augmented inverse probability weighted (AIPW) moment vector where the first term on the right hand side (RHS) of (6) is the IPW term, while the other terms on the RHS are the augmentation. Therefore, the weighted average representation in (3) implies that $\varphi_{[a,b]}(O;\beta)$ is also another AIPW moment vector, but concerning a different set of moments.

Lemma 5 summarizes in the current context the idea behind the Narain (1951)-Horvitz and Thompson (1952)-Hajek (1971) IPW principle under the general MAR condition in (1). For each $\beta \in \mathcal{B}$, this IPW principle enables identification of $E[m(Z;\beta)|a \leq C \leq b]$ whose sample version is infeasible, based on a quantity whose sample version is feasible.

Lemma 5 *If $P(C = R|T_R) > 0$ almost surely T_R then the general MAR condition in* (1) *implies that $E[m(Z;\beta)|a \leq C \leq b] = E\left[I(C = R)\omega^{IPW}_{[a,b]} m(Z;\beta)\right]$ for each $\beta \in \mathcal{B}$ where:*

$$\omega^{IPW}_{[a,b]} := \frac{\displaystyle\sum_{j=a}^{b} P(C = j|T_j, C \geq j) \prod_{r=1}^{j-1} \left[1 - P\left(C = r|T_r, C \geq r\right)\right]}{\displaystyle\prod_{r=1}^{R-1} \left[1 - P\left(C = r|T_r, C \geq r\right)\right] P(a \leq C \leq b)} = \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \omega_{R,j}(T_{R-1})$$

and where $\omega_{R,j}(T_{R-1})$ is defined in (5)*, and indeed $\omega_{R,j}(T_{R-1}) = \omega^{IPW}_{[j,j]}$ for $j = 1, \ldots, R$.*

For brevity we used the convention that if $a = 1$ then $\prod_{r=1}^{a-1}(1 - P(C = r|T_r, C \geq r)) = 1$.

Lemma 5 gives the foundation for IPW estimation based on an estimator of $E[I(C = R)\omega^{\text{IPW}}_{[a,b]} m(Z;\beta)]$, namely,

$$\frac{1}{n}\sum_{i=1}^{n} I(C_i = R)\widehat{\omega}^{\text{IPW}}_{[a,b]} m(Z_i;\beta) \tag{14}$$

as the GMM sample moment vector, where $\widehat{\omega}^{\text{IPW}}_{[a,b]}$ is an estimator of $\omega^{\text{IPW}}_{[a,b]}$ obtained by replacing each conditional hazard by its parametric or nonparametric estimator. In this section, our discussion of variance adjustment and efficiency in the context of the information content of the general MAR condition in (1) will correspond to nonparametric estimation of $\omega^{\text{IPW}}_{[a,b]}$.

Proposition 6 *(i) The "limited information" efficient GMM estimator of $\beta^0_{[a,b]}$ based on the moment restrictions:*

$$E\left[I(C=R)\omega^{IPW}_{[a,b]}m(Z;\beta^0_{[a,b]})\right]=0 \tag{15}$$

where for each $r=a\ldots,R-1$ the $P(C=r|T_r,C\geq r)$'s in $\omega^{IPW}_{[a,b]}$ solve for the $p_r(T_r)$'s from:

$$E\left[I(C\geq r)\left\{I(C=r)-p_r(T_r)\right\}\ \mid T_r\right]=0 \quad \textit{almost surely } T_r, \tag{16}$$

has asymptotic variance equal to the inverse of the semiparametric information bound for $\beta^0_{[a,b]}$ under the "full information" contained jointly in the restrictions (15) and (16).

(ii) Furthermore, this asymptotic variance from (i) is equal to $\Omega^{-1}_{[a,b]}$ where $\Omega_{[a,b]}:=M'_{[a,b]}V^{-1}_{[a,b]}M_{[a,b]}$ is defined in the statement of Proposition 2.

Proposition 6(i) applies Theorem 1 of Ackerberg et al. (2014) to show that the "limited information" and "full information" (using their terminology) efficient GMM estimation of $\beta^0_{[a,b]}$ based on (15) and (16) are equivalent in terms of the asymptotic variance of the estimator of $\beta^0_{[a,b]}$. Concretely, this "limited information" estimator is the efficient GMM estimator based on the IPW GMM sample moment vector in (14), i.e,

$$\widehat{\beta}^{IPW}_{[a,b]}(W_n):=\arg\min_{\beta\in\mathcal{B}}\left(\frac{1}{n}\sum_{i=1}^{n}I(C_i=R)\widehat{\omega}^{\mathrm{IPW}}_{[a,b]}m(Z_i;\beta)\right)'W_n\left(\frac{1}{n}\sum_{i=1}^{n}I(C_i=R)\widehat{\omega}^{\mathrm{IPW}}_{[a,b]}m(Z_i;\beta)\right) \tag{17}$$

when W_n is a consistent estimator of the asymptotic variance of the moment vector, accounting for the estimation of the nuisance conditional hazards $P(C=r|T_r,C\geq r)$'s involved in $\omega^{\mathrm{IPW}}_{[a,b]}$. The equivalence in asymptotic variance in Proposition 6(i) holds because the conditional hazards $P(C=r|T_r,C\geq r)$'s that constitute $\omega^{\mathrm{IPW}}_{[a,b]}$ are "exactly identified" by (16).

Proposition 6(ii) shows that this asymptotic variance in Proposition 6(i) reaches the semiparametric efficiency bound that was obtained in Proposition 2 under the general MAR condition in (1) for $\beta^0_{[a,b]}$ defined by (2). Thus, in the spirit of Graham (2011), we say that the moment restrictions (15) and (16) exhaust all available information about $\beta^0_{[a,b]}$ under the general setup of Proposition 2. Similar results with $R = 2$ are known from Hirano et al. (2003), Chen et al. (2008), Graham (2011), etc. but the case of $R > 2$ will help us to get further insights into this result and the information content of the MAR assumption.

We spend the rest of this section discussing Proposition 6(ii) with the following remarks.

First, there are two different semiparametric efficiency bounds present in Proposition 6: in (i) it is the bound based on the system (15) and (16), whereas in (ii) it is the bound based on our general framework (1) and (2). The result on semiparametric efficiency in Newey (1994), Ackerberg et al. (2014), etc. of the "limited information" approach concerns the first efficiency bound, i.e., the result in Proposition 6(i). On the other hand, the second efficiency bound is traditionally established independently in this literature, albeit in simpler contexts. Graham (2011) established the equality of these two bounds when $R = 2$ and $a = 1, b = R$ and $d_m = d_\beta$; however, his result was based on the Brown and Newey (1998)-orthogonalization that is not applicable here if interest lies on sub-populations.

Second, we find the equality of the two efficiency bounds remarkable in the case of $R > 2$ considering how much information the general MAR condition in (1) has and how little of it is used by the moment restrictions (15) and (16) leading to the first efficiency bound. In fact, (1) does not play any direct role in Proposition 6. (1)'s only role would be in the background ensuring that (15) holds, and Proposition 6(i) takes (15) as given.[11] The general MAR condition in (1) has no role to play in (16) — it contains no information about the unknown parameters in (16) since these moment restrictions simply follow from the definition of the conditional hazards and thus the parameters involved there are what is variously called "nonparametrically identified", "exactly identified" or "locally just identified"; see Newey (1994), Ackerberg et al. (2014), Chen and Santos (2018), Chernozhukov et al. (2018), etc.

Third, we point out that an equivalence result like Proposition 6(ii) will not hold if the general MAR condition in (1) is strengthened. The limited or full information approach will "pay a price" in terms of efficiency for not considering the (strengthened) MAR condition.

[11] For (15) to hold, it only requires the part of MAR with $r = a, \ldots, R-1$. The part with $r = 1, \ldots, a-1$ is unused since only the $P(C = r | T_r, C \geq r)$'s for $r = a, \ldots, R-1$ appear in the weight $\omega^{IPW}_{[a,b]}$; see (5).

For a clean demonstration of paying a price, Lemma 7(ii) strengthens (1) by imposing an extreme dimension reduction on the conditioning set leading to the CMAR condition in (13).

Lemma 7 *(i) The efficient GMM estimator of $\beta^0_{[a,b]}$ based on the moment restrictions:*

$$E\left[\sum_{j=a}^{b}\frac{P(C=j)}{P(a\leq C\leq b)}\frac{I(C=R)}{\prod_{r=j}^{R-1}(1-P(C=r|T_1,C\geq r))}\frac{P(C=j|T_1,C\geq j)}{P(C=j)}m(Z;\beta^0_{[a,b]})\right]=0,$$

where for each $r=a\ldots,R-1$ the $P(C=r|T_1,C\geq r)$'s solve for the $p_r(T_1)$'s from:

$$E\left[I(C\geq r)\left\{I(C=r)-p_r(T_1)\right\}\ \mid T_1\right]=0\quad \textit{almost surely } T_1,$$

has the same asymptotic variance $\left(M'_{[a,b]}\left[V^{\dagger}_{[a,b]}\right]^{-1}M_{[a,b]}\right)^{-1}$ under both the "limited and full information" approaches under regularity conditions where $V^{\dagger}_{[a,b]}:=E\left[\varphi^{\dagger}_{[a,b]}\varphi^{\dagger\prime}_{[a,b]}\right]$ and:

$$\varphi^{\dagger}_{[a,b]}=\frac{I(C=R)}{P(C=R|T_1)}\frac{P(a\leq c\leq b|T_1)}{P(a\leq C\leq b)}\left[m(Z;\beta^0_{[a,b]})-E[m(Z;\beta^0_{[a,b]})|T_1]\right]+\frac{I(a\leq C\leq b)}{P(a\leq C\leq b|T_1)}E[m(Z;\beta^0_{[a,b]})|T_1].$$

(ii) The inverse of the semiparametric information bound for $\beta^0_{[a,b]}$ in Proposition 4 that works under the CMAR condition in (13) cannot exceed this asymptotic variance $\left(M'_{[a,b]}\left[V^{\dagger}_{[a,b]}\right]^{-1}M_{[a,b]}\right)^{-1}$ because $V^{\dagger}_{[a,b]}-V^{CMAR}_{[a,b]}$ is positive semi-definite since $V^{\dagger}_{[a,b]}-V^{CMAR}_{[a,b]}$ is given by:

$$\sum_{r=2}^{R}E\left[\frac{P(a\leq C\leq b|T_1)}{P(a\leq C\leq b)}\left\{\frac{1}{P(C\geq R|T_1)}-\frac{1}{P(C\geq r|T_1)}\right\}Var\left(E[m(Z;\beta^0_{[a,b]})|T_r]\mid T_{r-1}\right)\middle|\,a\leq C\leq b\right].$$

The moment function for $\beta^0_{[a,b]}$ in Lemma 7(i) could be more compactly written as the weighted average of the $I(C=R)\omega_{R,j}(T_1)m(Z;\beta^0_{[a,b]})$'s where $\omega_{R,j}(.)$ is defined in (5). However, the more elaborate form in the lemma helps to better visualize where/how the variance adjustment, as in Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2018), etc., works in this IPW estimation. It works in Lemma 7(i) because there the conditional hazards are still exactly identified by the respective conditional moment restrictions. By contrast, the variance adjustment of IPW does not (and should not) work in Lemma 7(ii) in a way that leads to the efficient influence function and efficiency bound from Proposition 4.

The variance adjustment of IPW does not lead to the efficiency bound from Proposition 4 because the variance adjustment is based only on the given moment restrictions and no other information such as the MAR or CMAR conditions.[12] Note, e.g., that the CMAR condition in (13) contains additional information $P(C=r|Z,C\geq r)=\ldots=P(C=r|T_r,C\geq r)=\ldots=P(C=r|T_1,C\geq r)$ about the nuisance conditional hazards $P(C=r|T_1,C\geq r)$ in Lemma 7(i), thus providing a sequence of additional feasible moment restrictions:

$$E\left[I(C\geq r)\left\{I(C=r)-p_r(T_1)\right\}\ \mid T_j\right]=0\ \text{ almost surely } T_j \text{ for } j=1,\ldots,r$$

to solve for the $p_r(T_1)$'s in Lemma 7(i). Lemma 7(i) does not use this additional information and hence the IPW variance adjustment fails to reach the efficiency bound in Proposition 4.

While CMAR is a strong assumption, other types of strengthening of MAR — e.g., $P(C=r|Z,C\geq r)=P(C=r|Z_r,C\geq r)$, i.e, with conditioning set involving only period r's observables and not the entire history — could be more plausible. In general, the common empirical practice of any kind of variable selection also leads to an implicit strengthening of the MAR condition by imposing exclusion restrictions. It is likely that in such cases the IPW variance adjustment will also not lead to the efficiency bound like in the CMAR example.

Finally, we note that despite Proposition 6 and the elegant theory in the literature behind the variance adjustment of IPW estimators based on nonparametric estimation of the conditional hazards, IPW is not our recommended estimator even under MAR. The theory depends crucially on proper conditioning on the conditioning sets T_r's. However, the dimension of the conditioning set T_r increases with r, and in practice it is difficult to condition on all those variables especially if they are continuous. This makes the theory of nonparametric

[12] This did not matter in Proposition 6(ii) that worked under the MAR condition (1) because MAR did not have any information on the nuisance conditional hazards $P(C=r|T_r,C\geq r)$ in Proposition 6(i). MAR's information $P(C=r|Z,C\geq r)=P(C=r|T_r,C\geq r)$ cannot be feasibly used based on the observed data for efficiency via over identification of $P(C=r|T_r,C\geq r)$. This led to the equivalence in Proposition 6(ii).

variance adjustment less reflective of the finite-sample behavior even in very large samples when $R > 2$, which is a key feature of our paper. Simulations in Supplemental Appendix B suggest that nonparametric variance adjustment can underestimate IPW's true variability (measured by Monte Carlo variance) even in very large samples when $R > 2$, while parametric variance adjustment in (17) in the spirit of Newey (1994) or Ackerberg et al. (2012) can reflect the true variability in moderately large samples. This issue with IPW is distinct from the problems with IPW (primarily concerning bias) that have been noted in the recent double robustness literature; see Rothe and Firpo (2019), Chernozhukov et al. (2018), etc.

4 Estimator of $\beta^0_{[a,b]}$ and its asymptotic properties

Our proposed estimator for $\beta^0_{[a,b]}$ will utilize the doubly robust structure of $\varphi_{[a,b]}(O;\beta)$ that was highlighted in remark five following Proposition 2. We know from (3), (4) and (5) that $\varphi_{[a,b]}(O;\beta)$ depends on the unknown conditional hazards and conditional expectations. Denote the true value of these nuisance parameters by $p^0(T_{R-1})$ and $q^0(T_{R-1};\beta)$ where:

$$\begin{aligned} p^0(T_{R-1}) &:= \ (P(C = R-1|T_{R-1}, C \geq R-1), \ldots, P(C = a|T_a, C \geq a))', \\ q^0(T_{R-1};\beta) &:= (E[m(Z;\beta)|T_{R-1}]', \ldots, E[m(Z;\beta)|T_a]')'. \end{aligned}$$

Let $p(T_{R-1})$ and $q(T_{R-1};\beta)$ be generic functions of same dimension as $p^0(T_{R-1})$ and $q^0(T_{R-1};\beta)$.

Define the function $g(O;\beta,p(T_{R-1}),q(T_{R-1};\beta))$ as $\varphi_{[a,b]}(O;\beta)$ with the conditional hazards and conditional expectations replaced by the concerned elements of $p(T_{R-1})$ and $q(T_{R-1};\beta)$ respectively. Note that $g(O;\beta,p^0(T_{R-1}),q^0(T_{R-1};\beta)) \equiv \varphi_{[a,b]}(O;\beta)$ for all β.

We will use the following $d_m \times 1$ GMM sample moment vector to estimate the $d_\beta \times 1$ $\beta^0_{[a,b]}$:

$$\bar{g}_n(\beta,\widehat{p}(T_{R-1}),\widehat{q}(T_{R-1},\beta)) := \frac{1}{n}\sum_{i=1}^{n} g(O_i;\beta,\widehat{p}(T_{R-1,i}),\widehat{q}(T_{R-1,i},\beta)) \qquad (18)$$

where $\widehat{p}(T_{R-1,i})$ and $\widehat{q}(T_{R-1,i},\beta)$ are nonparametric or parametric estimators of $p^0(T_{R-1,i})$ and $q^0(T_{R-1,i};\beta)$ for $i = 1,\ldots,n$; see Robins and Rotnitzky (1992), Robins et al. (1994), etc.

For a $d_m \times d_m$ weighting matrix W_n, we will define the GMM estimator of $\beta^0_{[a,b]}$ as:

$$\widehat{\beta}(W_n) := \arg\min_{\beta\in\mathcal{B}} \left[\bar{g}_n(\beta, \widehat{p}(T_{R-1}), \widehat{q}(T_{R-1}, \beta))\right]' W_n \left[\bar{g}_n(\beta, \widehat{p}(T_{R-1}), \widehat{q}(T_{R-1}, \beta))\right]. \qquad (19)$$

Practitioners often use flexible parametric models to estimate the nuisance parameters. If there is “promise” to make the models more flexible when sample size increases then such estimators can be considered as nonparametric, otherwise they are parametric; see, e.g., Newey (1994) (p. 1369), Ackerberg et al. (2012), etc. We adopt this convention in our paper and provide a brief heuristic discussion of the properties of $\widehat{\beta}(W_n)$ by considering both parametric and nonparametric estimation of the nuisance parameters under a unified framework. Some generality is lost due to the unified presentation; but these results are already well known and our presentation here is only for the sake of completeness.

First, consider the conditional hazards. Let the parametric model, e.g., logit/probit, for $P(C = r|T_r, C \geq r)$ be $p_r(T_r; \gamma_r)$ where γ_r is a $d_{\gamma_r} \times 1$ unknown vector for $r = a, \ldots, R-1$. We obtain the quasi-maximum likelihood estimator $\widehat{\gamma}_r$ of γ_r solving the score equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} S_r(O_i; \widehat{\gamma}_r) \quad \text{for} \quad r = a, \ldots, R-1, \quad \text{where for} \quad i = 1, \ldots, n, \\ S_r(O_i; \gamma_r) &:= I(C_i \geq r)\frac{I(C_i = r) - p_r(T_{r,i}; \gamma_r)}{p_r(T_{r,i}; \gamma_r)(1 - p_r(T_{r,i}; \gamma_r))}\left\{\frac{\partial}{\partial\gamma_r} p_r(T_{r,i}; \gamma_r)\right\}. \end{aligned} \qquad (20)$$

Now, consider the conditional expectations. Let the parametric model, e.g., linear model, for the j-th element $E[m_j(Z; \beta)|T_r]$ of $E[m(Z; \beta)|T_r]$ be $q_{r,j}(T_r; \beta, \lambda_{r,j}(\beta))$. Let $q_r(T_r; \beta, \lambda_r(\beta)) = (q_{r,1}(T_r; \beta, \lambda_{r,1}(\beta)), \ldots, q_{r,d_m}(T_r; \beta, \lambda_{r,d_m}(\beta)))'$ where $\lambda_r(\beta) = (\lambda'_{r,1}(\beta), \ldots, \lambda'_{r,d_m}(\beta))'$ and $\lambda_{r,j}(\beta)$ is a $d_{\lambda_{r,j}} \times 1$ unknown vector for $r = a, \ldots, R-1$, $j = 1, \ldots, d_m$. We obtain the least squares estimator $\widehat{\lambda}_{r,j}(\beta)$ of $\lambda_{r,j}(\beta)$ for $j = 1, \ldots, d_m$ as functions of β solving the normal equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} L_{r,j}(O_i; \beta, \widehat{\lambda}_{r,j}(\beta)) \quad \text{for} \quad r = a, \ldots, R-1, \quad \text{where for} \quad i = 1, \ldots, n, \\ L_{r,j}(O_i; \beta, \lambda_r) &:= I(C_i = R)\left\{\frac{\partial}{\partial\lambda_{r,j}} q_{r,j}(T_{r,i}; \beta, \lambda_{r,j})\right\}(m_j(T_{R,i}; \beta) - q_{r,j}(T_{r,i}; \beta, \lambda_{r,j})). \end{aligned} \qquad (21)$$

In empirical work, the $p_r(T_r; \gamma_r)$'s are typically logit/probit with index $\xi'_{d_{\gamma_r}}(T_r)\gamma_r$, and the

$q_{r,j}(T_r;\beta,\lambda_{r,j}(\beta))$'s are typically linear $\pi'_{d_{\lambda_{r,j}}}(T_r)\lambda_{r,j}(\beta)$ where the $\xi_{d_{\gamma_r}}(T_r)$'s and $\pi_{d_{\lambda_{r,j}}}(T_r)$'s are possibly the first d_{γ_r} and $d_{\lambda_{r,j}}$ terms of some basis function; e.g., powers. We consider the estimator $\widehat{p}(T_{R-1}) = (p_{R-1}(T_{R-1};\widehat{\gamma}_{R-1}),\ldots,p_a(T_a;\widehat{\gamma}_a))'$ for $p^0(T_{R-1})$ and the estimator $\widehat{q}(T_{R-1};\beta) = (q'_{R-1}(T_{R-1};\beta,\widehat{\lambda}_{R-1}(\beta)),\ldots,q'_a(T_a;\beta,\widehat{\lambda}_a(\beta)))'$ for $q^0(T_{R-1};\beta)$ as parametric if the d_{γ_r}'s and $d_{\lambda_{r,j}}$'s are fixed, and as nonparametric if the d_{γ_r}'s and $d_{\lambda_{r,j}}$'s increase with n.

Assumption CH: The conditional hazard (CH) models are correct, i.e., there exists a $\gamma^0 = (\gamma_a^{0\prime},\ldots,\gamma_{R-1}^{0\prime})'$ such that $p_r(T_r;\gamma_r^0) = P(C = r|T_r, C \geq r)$ for $r = a,\ldots,R-1$.

Assumption CE: The conditional expectation (CE) models are correct, i.e., there exists a $\lambda^0 = (\lambda_a^{0\prime},\ldots,\lambda_{R-1}^{0\prime})'$ such that $q_r(T_r;\beta^0_{[a,b]},\lambda^0_r) = E[m(Z;\beta^0_{[a,b]})|T_r]$ for $r = a,\ldots,R-1$.

Assumptions CH and CE can be assumed to hold approximately arbitrarily well if $\widehat{p}(T_{R-1})$ and $\widehat{q}(T_{R-1};\beta)$ are nonparametric. But assumptions CH and CE may not hold if $\widehat{p}(T_{R-1})$ and $\widehat{q}(T_{R-1};\beta)$ are parametric. We will assume that $\|\widehat{p}-p^*\| = o_p(1)$ and $\|\widehat{q}-q^*\| = o_p(1)$ (at suitable rates and with respect to suitable metrics in suitable function spaces) for some pseudo true functions $p^*(T_{R-1})$ and $q^*(T_{R-1};\beta)$ where $p^*(T_{R-1}) = p^0(T_{R-1})$ if CH holds and $q^*(T_{R-1};\beta^0_{[a,b]}) = q^0(T_{R-1};\beta^0_{[a,b]})$ if CE holds. If both CH and CE fail to hold then there is no protection of double robustness and the GMM moment for $\beta^0_{[a,b]}$ may be misspecified. Then, in case of over identification ($d_m > d_\beta$) there may be no solution to the GMM population moment restriction and the probability limit of $\widehat{\beta}(W_n)$, if it exists, may depend on the limiting behavior of W_n; see, e.g., Hall and Inoue (2003). Such probability limits may not be of interest in the related empirical literature where the focus is on the true value $\beta^0_{[a,b]}$ and not the pseudo true values. Therefore, in our heuristic discussion below of the asymptotic properties of $\widehat{\beta}(W_n)$, we will maintain that assumptions CH and CE cannot be jointly false.

First, consistency. Double robustness implies (see remark 5 following Proposition 2) that:

$$E[g(O;\beta,p^*(T_{R-1}),q^0(T_{R-1};\beta))] = E[g(O;\beta,p^0(T_{R-1}),q^*(T_{R-1};\beta))] = E[m(Z;\beta)|a \leq C \leq b].$$

Therefore, consistency $\widehat{\beta}(W_n) \xrightarrow{p} \beta^0_{[a,b]}$ follows under standard conditions (see, e.g., Theorem 1 of Chen et al. (2003)) if CH and CE are not jointly false.

Now, the asymptotic distribution of $\widehat{\beta}(W_n)$. We can see that the same double robustness property also implies that the $M_{[a,b]}$ defined in assumption A3 satisfies:

$$M_{[a,b]} = \frac{\partial}{\partial \beta'} E[g(O;\beta^0_{[a,b]}, p^0(T_{R-1}), q^*(T_{R-1};\beta))] = \frac{\partial}{\partial \beta'} E[g(O;\beta^0_{[a,b]}, p^*(T_{R-1}), q^0(T_{R-1};\beta^0_{[a,b]}))].$$

Let $G_p(\beta,p,q)[v_p]$ and $G_q(\beta,p,q)[v_q]$ be the pathwise derivatives of $E[g(O;\beta,p,q)]$ at p and q in the directions v_p and v_q such that $p+\tau v_p$ and $q+\tau v_q$ for $\tau \in [0,1]$ belong in the respective function spaces. We can see that the same double robustness property also implies that:

$$G_p(\beta^0_{[a,b]}, p^*(T_{R-1}), q^0(T_{R-1},\beta^0_{[a,b]})) = 0 \quad \text{and} \quad G_q(\beta^0_{[a,b]}, p^0(T_{R-1}), q^*(T_{R-1},\beta)) = 0. \tag{22}$$

Let $W_n \xrightarrow{p} W$. If $\widehat{\beta}(W_n) \xrightarrow{p} \beta^0_{[a,b]}$ as we just noted above, then it now follows under standard conditions (see, e.g., Theorem 2 of Chen et al. (2003)) that:

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) &= -\left(M'WM\right)^{-1} M'W\sqrt{n}\,\big[\bar{g}_n(\beta^0,p^*,q^*) \\ &\qquad + G_p(\beta^0,p^*,q^*)[\widehat{p}-p^*] + G_q(\beta^0,p^*,q^*)[\widehat{q}-q^*]\big] + o_p(1)\end{aligned}$$

writing the triple $\beta, p(T_{R-1}), q(T_{R-1};\beta)$ as β, p, q, and dropping the subscript $[a,b]$ for brevity.

Therefore, if assumption CH holds, then $p^*(T_{R-1}) = p^0(T_{R-1})$ and hence by (22):

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\left(M'WM\right)^{-1} M'W\sqrt{n}\,\big[\bar{g}_n(\beta^0,p^0,q^*) + G_p(\beta^0,p^*,q^*)[\widehat{p}-p^0]\big] + o_p(1).$$

So the estimation of the unknown conditional expectations $E[m(Z;\beta^0)|T_r]$'s has no effect on the asymptotic distribution of $\widehat{\beta}(W_n)$ if the conditional hazard models are correct.

Similarly, if assumption CE holds, then $q^*(T_{R-1};\beta^0) = q^0(T_{R-1};\beta^0)$ and hence by (22):

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\left(M'WM\right)^{-1} M'W\sqrt{n}\,\big[\bar{g}_n(\beta^0,p^*,q^0) + G_q(\beta^0,p^0,q^0)[\widehat{q}-q^0]\big] + o_p(1).$$

So the estimation of the unknown conditional hazards $P(C=r|T_r;C \geq r)$'s has no effect on the asymptotic distribution of $\widehat{\beta}(W_n)$ if the conditional expectation models are correct.

Finally, if both assumptions CH and CE hold, then we have $p^*(T_{R-1}) = p^0(T_{R-1})$ and

$q^*(T_{R-1};\beta^0) = q^0(T_{R-1};\beta^0)$ and hence by (22):

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\left(M'WM\right)^{-1} M'W\sqrt{n}\bar{g}_n(\beta^0, p^0, q^0) + o_p(1).$$

Now consider efficiency in the sense of Proposition 2. If $W^{-1} = Var\left(g(O;\beta^0, p^*(T_{R-1}), q^*(T_{R-1},\beta^0))\right)$ $=: V(\beta^0, p^*, q^*)$, which when CH and CE hold jointly is denoted by $V(\beta^0, p^0, q^0)$, then:

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\left(M'[V(\beta^0, p^0, q^0)]^{-1}M\right)^{-1} M'[V(\beta^0, p^0, q^0)]^{-1}\sqrt{n}\bar{g}_n(\beta^0, p^0, q^0) + o_p(1)$$

when CH and CE hold jointly. Now, since the moment vector $g(O;\beta, p, q)$ was defined such that $g(O;\beta^0, p^0, q^0) \equiv \varphi_{[a,b]}(O;\beta^0)$ (and hence $V(\beta^0, p^0, q^0) \equiv V_{[a,b]}$), it follows that:

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\Omega_{[a,b]}^{-1} M'_{[a,b]} V_{[a,b]}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{[a,b]}(O_i;\beta^0) + o_p(1) \tag{23}$$

where the non-$o_p(1)$ term on the RHS is the influence function from Proposition 2 which was shown to be efficient for any $[a,b]$ when $d_m = d_\beta$ and for $a = b$ or $a = 1, b = R$ when $d_m > d_\beta$. Under the conditions maintained in Proposition 2, it follows from (23) that:

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) \xrightarrow{d} N\left(0, \Omega_{[a,b]}^{-1}\right).$$

The related literature on the doubly or locally robust moment functions using nonparametric $\widehat{p}$ and $\widehat{q}$, or even parametric $\widehat{p}$ and $\widehat{q}$ but without allowing for the violation of CH or CE, focuses solely on (23) and takes $\Omega_{[a,b]}^{-1}$ as the asymptotic variance of $\widehat{\beta}(W_n)$ when $W_n \xrightarrow{p} V_{[a,b]}^{-1}$.

However, assumption CH or CE may not hold if $\widehat{p}$ and $\widehat{q}$ are parametric. Then the above asymptotically linear representations of $\widehat{\beta}(W_n)$ are not practically useful to obtain the asymptotic variance of $\widehat{\beta}(W_n)$ without more structure on $\widehat{p}$ and $\widehat{q}$. The usual solution is to exploit the parametric structure of $\widehat{p}$ and $\widehat{q}$, and obtain the asymptotic variance of $\widehat{\beta}(W_n)$ based on the standard stacked representation of the moment vectors for β, $\gamma := (\gamma'_a, \ldots, \gamma'_{R-1})'$ and $\lambda := (\lambda'_a, \ldots, \lambda'_{R-1})'$ where $\lambda_r := (\lambda'_{r,1}, \ldots, \lambda'_{r,d_m})'$ for $r = a, \ldots, R-1$.

Accordingly, consider the $\left(d_m + \sum_{r=a}^{R-1} d_{\gamma_r} + \sum_{r=a}^{R-1}\sum_{j=1}^{d_m} d_{\lambda_{r,j}}\right) \times 1$ stacked moment vector:

$$\psi(O_i;\beta,\gamma,\lambda) := \begin{bmatrix} g(O_i;\beta,p(Z_i;\gamma),q(Z_i;\beta,\lambda)) \\ \underline{S}(O_i;\gamma) \\ \underline{L}(O_i;\beta,\lambda) \end{bmatrix} \quad \text{where} \quad \underline{S}(O_i;\gamma) := \begin{bmatrix} S_a(O_i;\gamma_a) \\ \vdots \\ S_{R-1}(O_i;\gamma_{R-1}) \end{bmatrix},$$

$$\underline{L}(O_i;\beta,\lambda) := \begin{bmatrix} \underline{L}_a(O_i;\beta,\lambda_a) \\ \vdots \\ \underline{L}_{R-1}(O_i;\beta,\lambda_{R-1}) \end{bmatrix}, \quad \text{and} \quad \underline{L}_r(O_i;\beta,\lambda_r) := \begin{bmatrix} L_{r,1}(O_i;\beta,\lambda_{r,1}) \\ \vdots \\ L_{r,d_m}(O_i;\beta,\lambda_{r,d_m}) \end{bmatrix}$$

for $r = a, \ldots, R-1$. We will obtain the GMM estimator $\widehat{\beta}$ using the usual two-step GMM.

We will refer to $\widehat{\beta}$ as EFF (as in efficient). In step one, we use the identity matrix as the GMM weighting matrix to obtain the first step estimators $\bar{\beta}, \bar{\gamma}$ and $\bar{\lambda}$ for β, γ and λ, and estimate the efficient weighting matrix as $\widehat{\Sigma}_n^{-1}(\bar{\beta},\bar{\gamma},\bar{\lambda})$ where $\widehat{\Sigma}_n(\beta,\gamma,\lambda) := \sum_{i=1}^n \psi(O_i;\beta,\gamma,\lambda)\psi'(O_i;\beta,\gamma,\lambda)/n$. Step one is not needed if $d_m = d_\beta$. In step two, we obtain the efficient GMM estimators $\widehat{\beta}, \widehat{\gamma}$ and $\widehat{\lambda}$ by minimizing with respect to β, γ, λ the GMM objective function based on the efficient weighting matrix. Finally, we estimate the asymptotic variance of $\widehat{\beta}$, i.e., EFF, as the first $d_\beta \times d_\beta$ block diagonal of the GMM asymptotic variance matrix $\left(\widehat{\Psi}_n'(\widehat{\beta},\widehat{\gamma},\widehat{\lambda})\widehat{\Sigma}_n^{-1}(\widehat{\beta},\widehat{\gamma},\widehat{\lambda})\widehat{\Psi}_n(\widehat{\beta},\widehat{\gamma},\widehat{\lambda})\right)^{-1}$ where $\widehat{\Psi}_n(\widehat{\beta},\widehat{\gamma},\widehat{\lambda})$ is the (possibly numerical) derivative of $\sum_{j=1}^n \psi(O_j;\beta,\gamma,\lambda)/n$ with respect to β, γ and λ at $\widehat{\beta}, \widehat{\gamma}$ and $\widehat{\lambda}$.

The asymptotic theory for EFF with parametric (fixed) nuisance models is simple. When CH and CE are not jointly false, the interesting structure described in the text between equations (22) and (23) is preserved by the influence function of EFF (and hence its asymptotic variance) thanks to the double robustness to the misspecification of the parametric nuisance models. If the parametric nuisance models are not fixed but "promise" to become sufficiently flexible with the increase in sample size, then, as shown in Ackerberg et al. (2012) (also see Newey (1994)), EFF can be interpreted as semiparametric and the estimator of its asymptotic variance obtained above can be consistent for the benchmark variance $\Omega_{[a,b]}^{-1}$.

5 Empirical Illustration based on Project STAR

We continue with the motivating example from Section 2.1 of attrition in the Project STAR. We wish to illustrate the possible benefits of the efficiency gains due to our proposed estimator EFF in drawing substantive conclusion from this experiment on the effect of small class size on students' performance. As a reference to EFF, we also present the same results using the IPW estimator from (14) that is the re-weighted Hajek (1971) version of IPW.

To this end, it is useful to first ask which effects an "ideal" Project STAR experiment would have generated with the subjects/students entering grade K in 1985 if there was no subsequent attrition or other implementation-related compromises; see, e.g., Hanushek (1999). The answer is that, since there was no protocol to randomly assign the class types of students except at the beginning, an "ideal" Project STAR experiment would have generated in grades K, 1, 2 and 3 the effect of continued presence in small classes with respect to continued presence in not-small classes. Our illustration will focus on the "ideal" experiment.

We first formally define these effects that the "ideal" experiment would have generated. We view attrition — a compromise to the ideal experiment — as a mitigating action by students in response to the treatment (class type) that they perceived as unhelpful to them. To gain a better understanding of this mitigating action we then decompose these effects by the attrition behavior of students from small and not-small classes.

For brevity of this illustration, we present only the results for (normalized) reading scores.[13] Let $Y^{\mathrm{s}}(\text{grade } j \text{ read})$ be the potential grade j reading score of a student *had* (s)he stayed in the small class at least until the end of grade j for $j = K, 1, 2, 3$ after being initially randomized to a small class in grade K. Similarly, with superscript "ns" denoting not-small, define the potential scores $Y^{\mathrm{ns}}(\text{grade } j \text{ read})$ for $j = K, 1, 2, 3$. These scores are not observed for a student in grade j if the student left the participating school before grade $j = 1, 2, 3$.

[13] To streamline our empirical illustration we ignore the compromises other than attrition to the experiment, e.g., students who enrolled after grade K or the few students (1.8%–5.8% in the respective grades; see Table 1) who switched their assigned class types. Some of these compromises can be accommodated in this illustration at the cost of strong modeling assumptions and messier notation that we want to avoid here for simplicity.

As noted above, we focus on two treatment regimes — a continued presence in small classes and a continued presence in not-small classes over the four years of Project STAR. Denote the average difference between the outcomes of these two regimes at each grade $j = K, 1, 2, 3$ as:

$$\mu_j^{\text{read}} := E[Y^{\text{s}}(\text{grade } j \text{ read}) - Y^{\text{ns}}(\text{grade } j \text{ read})].$$

Evolution of the effect of small classes:

First, consider the trajectory of μ_j^{read} for $j = K, 1, 2, 3$ to see how the effect of the small-class regime with respect to the not-small class regime evolved over continued presence in these regimes. Their EFF and IPW estimates are plotted in Figure 1(a).[14] The EFF and IPW estimates of the trajectory are quite similar. Consistent with the literature, we observe that the initial effect μ_K^{read} is very large compared to the "value added" (e.g., $\mu_j^{\text{read}} - \mu_K^{\text{read}}$ for $j = 1, 2, 3$) in the subsequent grades 1, 2 and 3. However, our value added estimates are not as pessimistic as Hanushek (1999)'s that led him to question the justification of the huge cost of prolonged operation of small classes, but are more in line with Krueger (1999).

We conjecture that the correction for attrition makes our estimates less pessimistic than Hanushek (1999)'s. This would happen under asymmetric selection, e.g., if the students leaving not-small classes left because they were going to score badly had they stayed whereas the students leaving small classes left under other concerns or lesser concerns of bad scores.

Following up on our conjecture, as proxies to Hanushek (1999)'s annual and 4-year sam-

[14] We obtain these estimates following Section 4 using parametric models specified for the conditional hazards and conditional expectations. The conditional hazard of leaving small (respectively, not-small) classes after grade j (= K, 1, 2) is modeled as logit with a linear index of a constant, dummies for race, sex, types of school (inner city, urban and rural), the share of students on free-lunch in school, dummies for all grades (present and past) where the student was on free lunch, where the student's teacher had bachelor's degree, and the difference in each of the past grades between the student's normalized math and reading scores from, respectively, the average normalized math and average normalized reading scores in small classes and also in not-small classes in their school. The differences between the student's and the average scores are continuous variables, and we also include their quadratic and cubic terms in the index. The conditional expectations of the grade j $(= 1, 2, 3)$ scores in small (respectively, not-small) classes are modeled linearly with exactly the same set of variables. These estimation results are not reported but are available from us.

ples respectively, we also plot in Figure 1(a) the "In grade" and "Never left" estimates of the trajectory. These are based on the average observed score of the students who took the tests at the end of the respective grades (for In grade estimates) and the students who continued in Project STAR until the end of grade 3 (for Never left estimates).[15] Note that, Never left actually estimates $\nu_{j,3}^{\text{read}}$ while In grade estimates $\nu_{j,j}^{\text{read}}$ for $j = K, 1, 2, 3$ where:

$$\nu_{j,l}^{\text{read}} \quad := \quad E[Y^{\text{s}}(\text{grade } j \text{ read})|B_l^{\text{s}}] \; - \; E[Y^{\text{ns}}(\text{grade } j \text{ read})|B_l^{\text{ns}}] \quad \text{for} \;\; j, l = K, 1, 2, 3$$

and B_l^{s} is the event that a student assigned to small class in grade K does not leave before the end of grade l for $l = K, 1, 2, 3$; and similarly B_l^{ns} is the event for the not-small class.[16]

Supporting our conjecture, visual inspection of In grade and Never left estimates reveals that without correction for attrition the value added estimates would indeed be pessimistic.

Does attrition matter?

But, beyond this visual inspection, does the correction for attrition matter statistically as well? More precisely, since we observed that the attrition-corrected estimates (EFF and IPW) are larger than the attrition-uncorrected estimates (In grade, which is typically favored to Never left), it is natural to ask if this is entirely due to sampling variation or is there systematic evidence for this in the population. That is, one would want to test the null hypothesis $H_{0,j} : \mu_j^{\text{read}} = \nu_{j,j}^{\text{read}}$ against the alternative $H_{1,j} : \mu_j^{\text{read}} > \nu_{j,j}^{\text{read}}$ for grades $j = 1, 2, 3$.

The p-values for these tests using EFF and IPW estimates of μ_j^{read} for grades $j = 1, 2, 3$

[15] "In grade" and "Never left" are those that correspond to the so-called "available cases" and "complete cases" respectively in the parlance of the missing data literature. To fix ideas consider Table 2. Never left has its own row in the table, while In grade for each grade is composed of the non-x entries in the column for that grade. In grade is preferred in practice (not always correctly) to Never left as a representative of the full population since it contains Never left and also units from various sub-populations of the full population.

[16] While we have deviated from the C-notation for attrition category to better reflect the sequencing $K, 1, 2, 3$ of grades, in this 4-period experiment: $B_K^{\text{s}} \equiv \{C \geq 1\}$ and $B_l^{\text{s}} \equiv \{C \geq l + 1\}$ if $l = 1, 2, 3$ for small class, and similarly B_l^{ns} for not-small class. We hope that this switch from C to B notation is not confusing.

Equipped with this notation, let us now recall and generalize the motivating discussion below Table 2 in Section 2.1 on the problem of selection. $\nu_{K,K}^{\text{read}} = \mu_K^{\text{read}}$ obviously as attrition started only after the end of grade K. However, in general $\nu_{j,l}^{\text{read}} \neq \mu_j^{\text{read}}$ for $j = K, 1, 2, 3$ and $l = 1, 2, 3$ unless suitable mean independence assumptions hold or, by happenstance, the biases for small and not-small classes cancel out, i.e., $E[Y^{\text{s}}(\text{grade } j \text{ read})|B_l^{\text{s}}] - E[Y^{\text{s}}(\text{grade } j \text{ read})] = E[Y^{\text{ns}}(\text{grade } j \text{ read})|B_l^{\text{ns}}] - E[Y^{\text{ns}}(\text{grade } j \text{ read})]$.

are as follows:

- 20.9% using EFF and 26.3% using IPW for $H_{0,1} : \mu_1^{\text{read}} = \nu_{1,1}^{\text{read}}$ against $H_{1,1} : \mu_1^{\text{read}} > \nu_{1,1}^{\text{read}}$. (Note that grade 1 score has a single level of missingness; see caption of Table 2.)
- 5.5% using EFF and 30.7% using IPW for $H_{0,2} : \mu_2^{\text{read}} = \nu_{2,2}^{\text{read}}$ against $H_{1,2} : \mu_2^{\text{read}} > \nu_{2,2}^{\text{read}}$.
- 6.6% using EFF and 23.3% using IPW for $H_{0,3} : \mu_3^{\text{read}} = \nu_{3,3}^{\text{read}}$ against $H_{1,3} : \mu_3^{\text{read}} > \nu_{3,3}^{\text{read}}$.

The EFF p-values for $H_{0,2}$ and $H_{0,3}$ are small and not sufficient in practice to take for granted the reliability of the attrition-uncorrected In grade estimates for the true effect μ_2^{read} and μ_3^{read}. On the other hand, the IPW p-values are quite a bit larger for $H_{0,2}$ and $H_{0,3}$. It is however not prudent (and possibly misleading) to take $H_{0,2}$ and $H_{0,3}$ for granted because, as we will see below, the large IPW p-values are entirely due to the imprecise nature of the IPW estimates. By contrast, EFF helps to avoid this possibly misleading confidence in $H_{0,2}$ and $H_{0,3}$ and points toward the possibility that attrition does matter here.

Do attrition-corrected estimates give substantive conclusions on the effects?

Attrition-correction will be of limited use to practitioners if it does not lead to precisely estimated (zero or non-zero) effects. To explore if that is the case here, we plot in Figure 1(b) the 90%, 95% and 99% two-sided confidence intervals around the EFF and IPW estimates for $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}$ and μ_3^{read}. The EFF intervals turn out to be subsets of the IPW intervals.

Specifically, while the EFF and IPW intervals are identical for μ_K^{read} by definition and are similarly precise for μ_1^{read} (one level of missingness), the EFF intervals are much more precise than the IPW intervals for μ_2^{read} and μ_3^{read} (more than one level of missingness).

EFF rejects a zero or negative value of μ_j^{read} for all $j = K, 1, 2, 3$ at all conventional levels but IPW fails to reject it for $j = 2, 3$ at the 1% level. (The EFF p-values do not exceed even .01%.) Small classes are an expensive policy proposition. Hence, the fact that EFF can rule out with extreme confidence any negative evidence against continued presence in small classes for every duration 1-4 years (after starting in grade K) has serious policy implications.

Attrition as a mitigating action against unhelpful class type assignment:

Students were randomly assigned to small and not-small classes when they enrolled in a Project STAR school in grade K. Many students did not score well in their randomly assigned class type. Leaving the Project STAR school was an important course of mitigating action available to these students. If attrition in Project STAR was primarily due to this mitigating action then, given the initial random assignment, we would expect that students who stayed scored better than what students who left would have scored had they stayed instead.

This is exactly what we observe in our estimates for each grade 1, 2 and 3. For brevity, we report here only the results for grade 3 since it is the terminal period of the experiment, and compare those who never left with each of the other attrition categories. Table 3 reports the EFF and IPW estimates of $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}}$ and $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}}$ for $j = K, 1, 2$ where:

$$\alpha_j^{\text{s,read}} \;:=\; E[Y^{\text{s}}(\text{grade 3 read}) \mid A_j^{\text{s}}], \quad \text{and} \quad \alpha_j^{\text{ns,read}} \;:=\; E[Y^{\text{ns}}(\text{grade 3 read}) \mid A_j^{\text{ns}}]$$

and A_j^{s} is the event that a student assigned to small class in grade K leaves exactly at the end of grade j; and similarly A_j^{ns} is the event for not-small classes.[17]

EFF and IPW estimates are very similar, but EFF is much more precise than IPW. Consequently, EFF confirms with a higher level of confidence in all cases the intuition that students who stayed scored better on average than what students who left would have scored had they stayed instead. By contrast, IPW fails to confirm at conventional levels of significance this intuition behind the choice to leave not-small classes at the end of grade 2.

Relatedly, consider the two decompositions of the effect μ_3^{read} of small classes by attrition categories:

$$\mu_3^{\text{read}} \;=\; \sum_{j=K,1,2,3} \mu_{3,j,*}^{\text{read}} \times P\left(A_j^{\text{s}}\right) \;=\; \sum_{j=K,1,2,3} \mu_{3,*,j}^{\text{read}} \times P\left(A_j^{\text{ns}}\right)$$

based on the attrition from small and not-small classes respectively, where for $j = K, 1, 2, 3$:

[17] This switch from the C to A notation in this 4-period experiment is trivial: $A_K^{\text{s}} \equiv \{C = 1\}$ and $A_j^{\text{s}} \equiv \{C = j+1\}$ if $j = 1, 2, 3$ for small class, and similarly A_j^{ns} for not-small class. As in footnote 16, this switch better reflects the sequencing $K, 1, 2, 3$ of grades and does so in small and not-small classes separately.

	$\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}}$		$\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}}$	
j	EFF	IPW	EFF	IPW
K	0.39***	0.34***	0.48***	0.48***
	(0.11)	(0.14)	(0.05)	(0.18)
1	0.45***	0.48**	0.64***	0.63***
	(0.16)	(0.24)	(0.08)	(0.19)
2	0.51***	0.47***	0.46***	0.46
	(0.11)	(0.20)	(0.09)	(0.53)

Table 3: EFF and IPW estimates and standard errors (in parentheses) for $\alpha_3^{\text{t,read}} - \alpha_j^{\text{t,read}}$ for $t = s, ns$ and $j = K, 1, 2$. *, ** and *** signify if the null that the parameter is zero is rejected against the alternative that it is greater than zero at the 10%, 5% and 1% level respectively.

$$\mu_{3,j,*}^{\text{read}} = E[Y^{\text{s}}(\text{grade } j \text{ read})|\ A_j^{\text{s}}] - E[Y^{\text{ns}}(\text{grade } j \text{ read})],$$
$$\mu_{3,*,j}^{\text{read}} = E[Y^{\text{s}}(\text{grade 3 read})] - E[Y^{\text{ns}}(\text{grade 3 read})|A_j^{\text{ns}}].$$

EFF and IPW estimates of these two decompositions, along with the 90%, 95% and 99% two-sided confidence intervals, are reported in Figures 1 (c)-(d) showing the relative contribution of each attrition category from small and not-small classes respectively toward the overall effect. Given the large number of students who left, it is important to understand what the effect would have been with respect to students leaving at various junctures of the experiment. $\mu_{3,*,j}^{\text{read}}$ and $\mu_{3,j,*}^{\text{read}}$ for $j = K, 1, 2, 3$ are those effects on the grade 3 reading scores.

Figure 1 (c) reveals that if we compare a randomly chosen student assigned to small class with a randomly chosen student assigned to not-small class who never left not-small class, then there is no benefit of small classes on the grade 3 reading score. The benefit on the grade 3 reading score is driven by the comparison of the former student with randomly chosen students assigned to not-small class who left not-small class after grade K, 1 or 2.

Figure 1 (d) reveals that if we compare a randomly chosen student assigned to not-small class with randomly chosen students assigned to small class who left small class after grade K or 1 or 2, then there is no harm to the grade 3 reading score due to not-small classes. The harm to the grade 3 reading score due to not-small classes is driven by the comparison of the former student with a randomly chosen student assigned to small class who never left small class. Thus, attrition was clearly a mitigating action against unhelpful class assignment.

These decompositions reveal such interesting patterns telling us which group of students (by attrition category) are driving the overall effect of small classes in the terminal period grade 3, and by how much. While the EFF estimates of the decompositions are very similar to the IPW estimates, the precision of EFF provides more statistical confidence toward confirming the contribution of each group of students to the overall effect of small classes.[18]

Lastly, as we noted in Section 2.1, these EFF-based inferences are precise mainly because the sub-population-specific components of the effects are estimated more precisely by EFF. Table 4 reports the results for EFF and IPW estimation of a subset of such components. Rows (a)-(c) correspond to the components marked with "x" in the columns for the grade 3 score in Table 2 that was presented in Section 2.1 as an empirical motivation behind the theoretical contribution of our paper. The gain in precision due to EFF is clear in all cases.

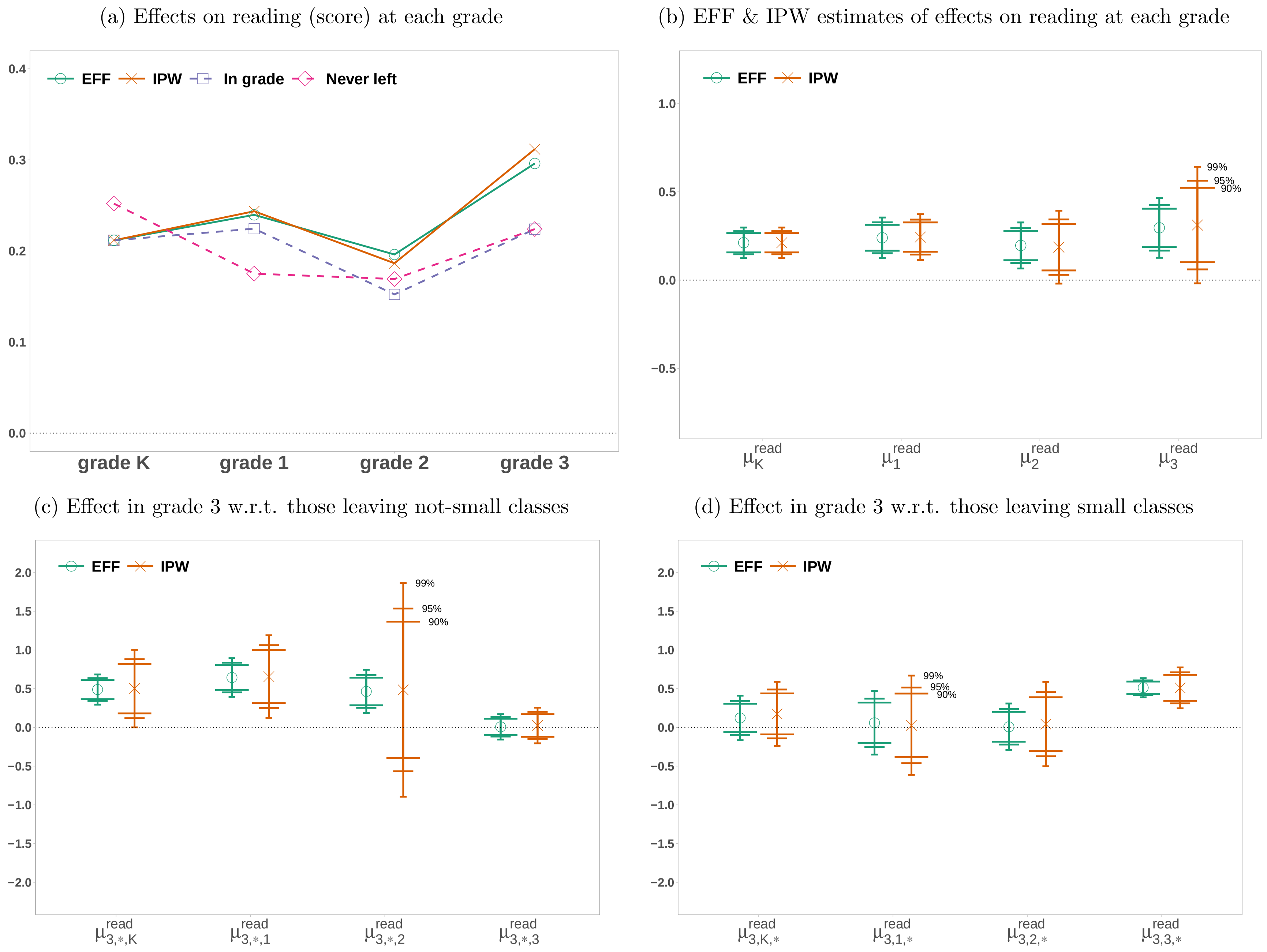
Left STAR school	Randomized to small class		Randomized to not-small class	
at the end of grade	EFF	IPW	EFF	IPW
(a) K	0.05	0.10	-0.27	-0.26
	(0.11)	(0.13)	(0.05)	(0.17)
(b) 1	-0.02	-0.04	-0.42	-0.42
	(0.16)	(0.23)	(0.08)	(0.19)
(c) 2	-0.07	-0.03	-0.24	-0.24
	(0.11)	(0.19)	(0.09)	(0.53)
(d) 3 (Never left)	0.44	0.44	0.22	0.22
	(0.04)	(0.04)	(0.03)	(0.03)

Table 4: EFF and IPW estimates of expected (counterfactual) reading scores in grade 3 by the student's attrition period are presented under the class types to which they were initially randomized. Standard deviations are presented in parentheses. All results in this empirical illustration are based on such parameters and the standard errors of those results were computed by noting that the estimates in this table across the two class types are independent but are correlated within class types. Row (d), i.e., Never left, involves nothing unobserved and hence both IPW and EFF estimates are equal to the simple group averages.

[18] Note that, for each $j = K, 1, 2$, the estimands from Table 3 are related to these decompositions as follows:

$$\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}} = \mu_{3,3,*}^{\text{read}} - \mu_{3,j,*}^{\text{read}} \quad \text{while} \quad \alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}} = -(\mu_{3,*,3}^{\text{read}} - \mu_{3,*,j}^{\text{read}}).$$

Therefore, going back to Table 3, we see that it suggests that EFF rejects the null $\mu_{3,*,3}^{\text{read}} = \mu_{3,*,j}^{\text{read}}$ against $\mu_{3,*,3}^{\text{read}} < \mu_{3,*,j}^{\text{read}}$ and the null $\mu_{3,3,*}^{\text{read}} = \mu_{3,j,*}^{\text{read}}$ against $\mu_{3,3,*}^{\text{read}} > \mu_{3,j,*}^{\text{read}}$ for each $j = K, 1, 2$ even at the 1% level. IPW cannot do that, and moreover it does not reject $\mu_{3,3,*}^{\text{read}} = \mu_{3,2,*}^{\text{read}}$ at any conventional level of significance.



(a) Effects on reading (score) at each grade

(b) EFF & IPW estimates of effects on reading at each grade

(c) Effect in grade 3 w.r.t. those leaving not-small classes

(d) Effect in grade 3 w.r.t. those leaving small classes

Figure 1: **(a)** EFF, IPW, In grade and Never left estimates of effect on reading score at each grade. **(b)** EFF and IPW estimates and confidence intervals (90%, 95%, 99%) of $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}, \mu_3^{\text{read}}$. 90%, 95%, 99% EFF and IPW confidence intervals for the decomposition of μ_3^{read} by comparing: **(c)** the entirety of small classes with different attrition categories from not-small classes, and **(d)** the different attrition categories from small classes with the entirety of not-small classes.

6 Conclusion

Our paper provided a comprehensive presentation of efficiency in estimation of parameters defined by the missingness pattern of monotonically missing at random data. The efficiency results on the parameters for generic sub-populations are new, and extend the well-known results on the treatment effects on the treated or the untreated or the parameters from the so-called “verify-out-of-sample” case in various empirically relevant directions.

We saw in the empirical illustration that such parameters are, among other things, fundamental to our understanding of the economic agent’s mitigation behavior when faced with unhelpful situations; e.g., leaving a school where a class-assignment is perhaps not working well for the student. Our proposed estimator for such parameters is a standard two-step doubly robust estimator. We saw that its computation is standard, and its precision may help to draw substantive conclusions when the standard estimators fail to do so. The excellent performance of our proposed estimator in our simulation experiment (Supplemental Appendix B) and, by contrast, the poor performance of its competitors give credibility to the results obtained by our proposed estimator; and we hope that encourages its use in practice.

We now conclude by recalling two important technical features of our paper.

First, we clearly characterized the additional restrictions that were imposed on the tangent set for the underlying semiparametric model by the over identification of the parameters of interest. To our knowledge, this characterization was missing from the related literature on missing data. In the process we validated and extended various existing results.

Second, we analyzed the information content (strength) of the MAR assumption linking it to the usability of sample units toward efficient estimation in sub-populations. This allowed us to contrast between the efficiency bound that is reached by the variance adjustment due to the estimation of exactly identified nuisance parameters and the efficiency bound that is obtained under the model assumptions involving the strength of the MAR assumption.

To our knowledge, these two technical features distinguish our paper from the related literature on missing data, and are possibly of independent interest for future work.

7 Bibliography

Abowd, J. M., Crepon, B., and Kramarz, F. (2001). Moment Estimation with Attrition: An Application to Economic Models. *Journal of the American Statistical Association*, 96:1223–1231.

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. *Review of Economics and Statistics*, 99:657–662.

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94:481–498.

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic Efficiency of Semiparametric Two-step GMM. *Review of Economic Studies*, 81: 919–943.

Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61:962–972.

Brown, B. and Newey, W. (1998). Efficient Semiparametric Estimation of Expectations. *Econometrica*, 66:453–464.

Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96:723–734.

Chaudhuri, S. (2020). On Efficiency Gains from Multiple Incomplete Subsamples. *Econometric Theory*, 36:488–525.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36:808–843.

Chen, X., Linton, O., and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criteria Function is not Smooth. *Econometrica*, 71:1591–1608.

Chen, X. and Santos, A. (2018). Overidentification in Regular Models. *Econometrica*, 86: 1771–1817.

Chernozhukov, V., Escanciano, J., Ichimura, H., Newey, W., and Robins, J. M. (2018). Locally Robust Semiparametric Estimation. Working Paper.

Chernozhukov, V., Escanciano, J.-C., Ichimura, H., Newey, W., and Robins, J. (2022). Locally Robust Ssemiparametric Estimation. *Econometrica*, 90: 1501–1535.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126:1593–1660.

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162:362–368.

Ding, W. and Lehrer, S. F. (2010). Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions. *The Review of Economics and Statistics*, 92:31–42.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1996). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper*. NBER.

Gill, R. and Robins, J. M. (1997). Non-Response Models For The Analysis Of Non-Monotone Ignorable Missing Data. *Statistics in Medicine*, 16:39–56.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures and Counterexamples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statitsics, pages 255–294. New York: Springer-Verlag.

Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79:437 – 452.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66:315–331.

Hajek, J. (1971). Comment on a paper by d. basu. In Godambe, V. R. and Sprott, D. A., editors, *Foundations of Statistical Inference*, page 236. Holt, Rinehert and Winston, Toronto.

Hall, A. R. and Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114: 361–394.

Hanushek, E. A. (1999). Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, 21:143–63.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores. *Econometrica*, 71:1161–1189.

Holcroft, C., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65:349–374.

Hoonhout, P. and Ridder, G. (2019). Nonignorable Attrition in Multi-Period Panels With Refreshment Samples. *Journal of Business and Economic Statistics*, 37:377–390.

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, 47:663–685.

Khan, S. and Tamer, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78:2021–2042.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 114:497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR. *The Economic Journal*, 111:1–28.

Muris, C. (2020). Efficient GMM estimation with incomplete data. *Review of Economics and Statistics*, 102: 518–530.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of Indian Soc. Agricultural Statistics*, 3:169–174.

Newey, W. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62:1349–1382.

Newey, W. K. (1990). Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, 5:99–135.

Nicoletti, C. (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132:461–489.

Robins, J. M. and Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16:39–56.

Robins, J. M. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models. *Statistics in Medicine*, 16:285–319.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In N. Jewell, K. D. and Farewell, V. T., editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331. Birkhliuser, Boston.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of American Statistical Association*, 90:122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of American Statistical Association*, 427:846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of American Statistical Association*, 429:106–121.

Rothe, C. and Firpo, S. (2019). Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. *Econometric Theory*, 35: 1048–1087.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika*, 82:805–820.

Rubin, D. (1976). Inference and Missing Data. *Biometrika*, 63:581–592.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22:560–568.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Vansteelandt, S., Rotnitzky, A., and Robins, J. M. (2007). Estimation of regression models for mean of repeated outcomes under nonignorable nonmotonone nonresponse. *Biometrika*, 94:841–860.

Wooldridge, J. M. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1:117–139.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section & Panel Data*. MIT Press.

Supplemental Appendix:

A note on efficiency in estimation with monotonically missing at random data

Jean-Louis Barnwell[19] and Saraswata Chaudhuri[20]

Table of Contents

The numbering of the equations, corollaries, lemmas and propositions in this supplemental appendix is consistent with the main text of our paper.

[19] Department of Economics, McGill University, Montreal, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

[20] Corresponding author. Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

A Supplemental Appendix A: Proofs

A.1 Auxiliary lemmas:

Lemma 8 *The MAR condition in* (1) *implies and is implied by the following condition:*

$$P(C = r|T_R) = P(C = r|T_r) \text{ for } r = 1, \ldots, R-1. \tag{24}$$

Proof of Lemma 8: First we show that if (1) holds then (24) also holds. Take any $r = 1, \ldots, R-1$ and note that:

$$\begin{aligned} P(C = r|T_R) &= P(C = r|T_R, C \geq r) \prod_{k=1}^{r-1} [1 - P(C = k|T_R, C \geq k)] \\ &= P(C = r|T_r, C \geq r) \prod_{k=1}^{r-1} [1 - P(C = k|T_k, C \geq k)] \quad \text{[by (1)]} \\ &= P(C = r|T_r, C \geq r) \prod_{k=1}^{r-1} [1 - P(C = k|T_r, C \geq k)] \quad \text{[by (1)]} \\ &= P(C = r|T_r). \end{aligned}$$

Now we show that if (24) holds then (1) also holds. Take any $r = 1, \ldots, R-1$ and note that:

$$\begin{aligned} P(C = r|T_R, C \geq r) &= \frac{P(C = r|T_R)}{P(C \geq r|T_R)} = \frac{P(C = r|T_R)}{1 - P(C \leq r-1|T_R)} = \frac{P(C = r|T_R)}{1 - \sum_{j=1}^{r-1} P(C = j|T_R)} \\ &= \frac{P(C = r|T_r)}{1 - \sum_{j=1}^{r-1} P(C = j|T_j)} = \frac{P(C = r|T_r)}{1 - \sum_{j=1}^{r-1} P(C = j|T_r)} \\ &= \frac{P(C = r|T_r)}{P(C \geq r|T_r)} = P(C = r|T_r, C \geq r) \end{aligned}$$

where the fourth and fifth equalities follow by (24). ■

Lemma 9 *The MAR condition in* (1) *implies that:*

$$P(C \geq r|T_j) = P(C \geq r|T_{r-1}) \text{ for } r = 1, \ldots, R-1 \text{ and } j = r, \ldots, R.$$

Proof of Lemma 9: Lemma 8 shows that (1) implies (24). Now, take $r = 1, \ldots, R-1$ and $j = r, \ldots, R$ and note that:

$$P(C \geq r|T_j) = 1 - \sum_{k=1}^{r-1} P(C = k|T_j) = 1 - \sum_{k=1}^{r-1} P(C = k|T_k) = 1 - \sum_{k=1}^{r-1} P(C = k|T_{r-1}) = P(C \geq r|T_{r-1})$$

where the second and third equalities follow by (24). ■

Remarks:

1. Lemma 9 implies that if $R = 2$ then $P(C = 2|T_2) = P(C = 2|T_1)$. This is the familiar form in which the MAR assumption is generally found in the econometrics literature where the focus has typically on the case of $R = 2$.

2. We introduced the notation in the above two lemmas for brevity of expressions in the proofs in this appendix. The original notation with the conditional hazards is very transparent in terms of accounting for the observability of the conditioning variables (and hence for estimation), and precisely for that reason it leads to longer expressions.

Lemma 10 *Under the conditions of Proposition 2 and using the notation of Section 3.2:*

$$E\left[\varphi_{[a,b]}(O, \beta^0_{[a,b]})S(O)'\right] = E\left[m \left\{ s(Z) + \frac{\dot{P}(a \leq C \leq b|T_b)}{P(a \leq C \leq b|T_b)} \right\}' \middle| a \leq C \leq b \right].$$

Proof of Lemma 10: Note from (3), (4) and (5) that:

$$\begin{aligned} \varphi_{[a,b]}(O; \beta^0_{[a,b]}) \quad = \quad & \sum_{r=b+1}^{R} \frac{I(C \geq r)}{P(C \geq r|T_{r-1})} \frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)} \left(E[m|T_r] - E[m|T_{r-1}]\right) \\ & + \sum_{r=a+1}^{b} \frac{I(C \geq r)}{P(C \geq r|T_{r-1})} \frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)} \left(E[m|T_r] - E[m|T_{r-1}]\right) \\ & + \sum_{r=a}^{b} \frac{I(C = r)}{P(a \leq C \leq b)} E[m|T_r]. \end{aligned} \qquad (25)$$

This alternative formulation of $\varphi_{[a,b]}(O; \beta^0_{[a,b]})$ without the conditional hazards stated explicitly is not intuitively transparent for actual computational purpose, but will be adopted here since it provides shorter expressions in the proof. Based on (25) we can write $E[\varphi_{[a,b]}(O;\beta^0_{[a,b]})S(O)'] = \sum_{i=1}^{3}\sum_{j=1}^{2} B_{ij}$ where:

$$
\begin{aligned}
B_{11} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)D'\right],\\
B_{12} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{21} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)D'\right],\\
B_{22} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq r-1|T_{r-1})}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\right.\\
&\qquad\qquad\qquad\left.\times\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{31} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]D'\right],\\
B_{32} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
D &:= s(Z_1)+\sum_{k=2}^{R} I(C\geq k)s(Z_k|T_{k-1}).
\end{aligned}
$$

We wrote this with the understanding that if $b=R$ then $B_{11}=B_{12}=0$, and if $a=b$ then $B_{21}=B_{22}=0$. For notational brevity define T_0 as any constant, so that $s(Z_1)\equiv s(Z_1|T_0)$. First, note that:

$$
\begin{aligned}
B_{11} &= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right]\\
&\quad+\sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{I(C\geq k)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right]\\
&= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{P(C\geq r|T_{r-1})}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right]\\
&\quad+\sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{P(C\geq k|T_{k-1})}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right]
\end{aligned}
$$

where the third and fourth lines follow by Lemma 9. Hence, we subsequently obtain that:

$$\begin{aligned}
B_{11} &= \sum_{r=b+1}^{R} E\left[\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)} E[m|T_r]s(Z_r|T_{r-1})'\right] + 0 \\
&= E\left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} ms(Z_R, \ldots, Z_{b+1}|T_b)'\right] \\
&= E\left[ms(Z_R, \ldots, Z_{b+1}|T_b)'|a \leq C \leq b\right]. \qquad (26)
\end{aligned}$$

The first equality follows since for all $k = 1, \ldots, r-1$: $E\left[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'\right] = E\left[E[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'|T_{r-1}]\right] = 0$ while for $k \geq r+1$: $E\left[E[m|T_r]s(Z_k|T_{k-1})'\right] = E\left[E[m|T_r]E[s(Z_k|T_{k-1})'|T_{k-1}]\right] = 0$. The second equality follows by (1) and Lemma 8 and the definition of score. The last equality is obvious.

Second, following the steps that led to the first line on the RHS of (26), we obtain that:

$$B_{21} = \sum_{r=a+1}^{b} E\left[\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)} E[m|T_{r-1}]s(Z_r|T_{r-1})'\right].$$

Therefore,

$$\begin{aligned}
B_{21} &= \sum_{r=a+1}^{b} \sum_{k=a}^{r-1} E\left[\frac{P(C = k|T_k)}{P(a \leq C \leq b)} ms(Z_r|T_{r-1})'\right] \\
&= \sum_{r=a+1}^{b} \sum_{k=a}^{r-1} E\left[ms(Z_r|T_{r-1})'|C = k\right] \frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[m \sum_{r=k+1}^{b} s(Z_r|T_{r-1})' \middle| C = k\right] \frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[ms(Z_b, \ldots, Z_{k+1}|T_k)'|C = k\right] \frac{P(C = k)}{P(a \leq C \leq b)}. \qquad (27)
\end{aligned}$$

The first equality follows by (1) and Lemma 8. The second equality follows by the same steps that gave the second line on the RHS of (26). The third equality follows by interchanging the order of summations (allowed here). The last equality follows by the definition of score.

Third, we consider B_{31} and note that using the definition of score in the second equality below and the same argument as before in the third (last) equality below give:

$$\begin{aligned}
B_{31} &= \sum_{r=a}^{b}\sum_{k=1}^{r} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(T_r)'\right] \\
&= \sum_{r=a}^{b} E\left[ms(T_r)'|C=r\right]\frac{P(C=r)}{P(a\leq C\leq b)}. \qquad (28)
\end{aligned}$$

Adding (27) and (28) gives:

$$\begin{aligned}
B_{21}+B_{31} &= E\left[ms(T_b)'|C=b\right]\frac{P(C=b)}{P(a\leq C\leq b)} + \sum_{k=a}^{b-1} E\left[ms(T_b)'|C=k\right]\frac{P(C=k)}{P(a\leq C\leq b)} \\
&= E\left[ms(T_b)'|a\leq C\leq b\right]. \qquad (29)
\end{aligned}$$

Now, we consider the terms B_{12}, B_{22} and B_{32} respectively. Accordingly, first note that:

$$\begin{aligned}
B_{12} &= \sum_{r=b+1}^{R}\sum_{k=r}^{R} E\left[\frac{I(C=k)}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right] \\
&= \sum_{r=b+1}^{R} E\left[\frac{1}{P(C\geq r|T_{r-1})}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=r}^{R}\dot{P}(C=k|T_k)'\right] \\
&= \sum_{r=b+1}^{R} E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C\geq r|T_{r-1})'}{P(C\geq r|T_{r-1})}\right] \\
&= 0. \qquad (30)
\end{aligned}$$

The second equality follows by (1) and Lemma 8. The third equality follows by Lemma 8 and Lemma 9. The fourth (last) equality follows by taking expectation conditional on T_{r-1} for the r-th term inside the summation. Exactly following the same steps as in the above (recall the analogy with B_{11} and B_{12} above) we obtain:

$$B_{22} = 0. \qquad (31)$$

Lastly, as before, note that:

$$
\begin{aligned}
B_{32} &= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(C=r|T_r)}\frac{E[m|T_r]\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right] = E\left[m\sum_{r=a}^{b}\frac{\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right]\\
&= E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b|T_b)}m\frac{\dot{P}(a\leq C\leq b|T_b)'}{P(a\leq C\leq b)}\right] = E\left[\frac{I(a\leq C\leq b)}{P(a\leq C\leq b|T_b)}m\frac{\dot{P}(a\leq C\leq b|T_b)'}{P(a\leq C\leq b)}\right]\\
&= E\left[m\frac{\dot{P}(a\leq C\leq b|T_b)'}{P(a\leq C\leq b|T_b)}\middle| a\leq C\leq b\right] \qquad (32)
\end{aligned}
$$

Therefore, (26) and (29)-(32) give the result. ■

A.2 Proof of the results stated in the main text

Proof of Lemma 1: For simplicity we suppress the dependence of quantities on O, Z, T_r, β, etc. unless confusing. Taking $a = 1, b = R$ in (3), note by using (4) and (5) that $\varphi_{[1,R]}(.)$ is:

$$
\begin{aligned}
&\sum_{j=1}^{R} P(C=j)\left\{\sum_{r=j+1}^{R}\frac{I(C\geq r)P(C=j|T_j)}{P(C=j)P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) + \frac{I(C=j)}{P(C=j)}E[m|T_j]\right\}\\
=& \sum_{r=2}^{R}\sum_{j=1}^{r-1} I(C\geq r)\frac{P(C=j|T_j)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) + \sum_{j=1}^{R} I(C=j)E[m|T_j]\\
=& \sum_{r=2}^{R} I(C\geq r)\frac{P(C\leq r-1|T_{r-1})}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) + \sum_{r=1}^{R} I(C=r)E[m|T_r]\\
=& \sum_{r=2}^{R} I(C\geq r)\frac{1-P(C\geq r|T_{r-1})}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) + \sum_{r=1}^{R} I(C=r)E[m|T_r]\\
=& \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) - \sum_{r=2}^{R} I(C\geq r)(E[m|T_r]-E[m|T_{r-1}]) + \sum_{r=1}^{R} I(C=r)E[m|T_r]\\
=& \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}]) + E[m|T_1]
\end{aligned}
$$

where the last line follows because we can write $\sum_{r=2}^{R} I(C\geq r)(E[m|T_r]-E[m|T_{r-1}])$ as:

$$
I(C=R)E[m|T_R] + \sum_{r=2}^{R-1} E[m|T_r][I(C\geq r)-I(C\geq r+1)] + I(C\geq 2)E[m|T_1].
$$

The law of iterated expectations gives $E[\varphi_{[1,R]}(O;\beta)] = 0 + E[E[m(Z;\beta)|T_1]]$. Therefore,

$$\begin{aligned}
E[\varphi_{[1,R]}] &= E\left[\varphi_{[1,R]}\left\{\varphi_{[1,R]} - E[E[m(Z;\beta)|T_1]]\right\}'\right] \\
&= E\left[\sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}])\sum_{s=2}^{R}\frac{I(C\geq s)}{P(C\geq s|T_{s-1})}(E[m|T_s]-E[m|T_{s-1}])'\right] \\
&\quad + E[E[m|T_1](E[m|T_1]-E[E[m|T_1]])'] \\
&= \sum_{r=2}^{R} E\left[\frac{1}{P(C\geq r|T_{r-1})}E\left[(E[m|T_r]-E[m|T_{r-1}])(E[m|T_r]-E[m|T_{r-1}])'|T_{r-1}\right]\right] \\
&\quad + E[E[m|T_1](E[m|T_1]-E[m])'] \\
&= \sum_{r=2}^{R} E\left[\frac{V(E[m|T_r]|T_{r-1})}{P(C\geq r|T_{r-1})}\right] + V(E[m|T_1])
\end{aligned}$$

giving the desired result, where the last equality follows simply by definition while the third equality follows since for $r > s$ by using (1) and the law of iterated expectations:

$$\begin{aligned}
& E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}])\frac{I(C\geq s)}{P(C\geq s|T_{s-1})}(E[m|T_s]-E[m|T_{s-1}])'\right] \\
= \; & E\left[\frac{1-I(C\leq r-1)}{(1-P(C\leq r-1|T_{r-1}))P(C\geq s|T_{s-1})}(E[m|T_r]-E[m|T_{r-1}])(E[m|T_s]-E[m|T_{s-1}])'\right] \\
= \; & 0
\end{aligned}$$

and for $r > 1$, again by using (1) and the law of iterated expectations,

$$E\left[\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}(E[m|T_r]-E[m|T_{r-1}])(E[m|T_1]-E[E[m|T_1]])'\right] = 0. \blacksquare$$

Remark: Unless confusing, we will write $m(Z;\beta^0_{[a,b]})$ simply as m for brevity in the sequel.

Proof of Proposition 2: We obtained the tangent set $\mathcal{T}$ in Section 3.2. In the just identified case the tangent set is given by (7) in the case of any generic $[a,b]$ with $a,b \in \{1,\ldots,R\}$ and $a \leq b$, while in the over identified case the tangent set is given by (7) and the additional restriction (10) if $a = 1, b = R$, and by (7) and the additional restriction (11) if $a = b$.

The rest of the proof will proceed as follows. We will show that $\varphi_{[a,b]}(O,\beta^0_{[a,b]})$ satisfies the pathwise derivative condition for any generic $[a,b]$ with $a,b \in \{1,\ldots,R\}$ and $a \leq b$ in the over identified case ($d_m \geq d_\beta$). Thus, this will obviously be satisfied in the just identified case ($d_m = d_\beta$). Then we will show that the concerned influence function obtained from $\varphi_{[a,b]}(O,\beta^0_{[a,b]})$ belongs in $\mathcal{T}$ in the over identified case if $a=1, b=R$ or if $a=b$, and belongs in $\mathcal{T}$ in the just identified case for any generic $[a,b]$ with $a,b \in \{1,\ldots,R\}$ and $a \leq b$.

Taking any A that is a full row rank $d_\beta \times d_m$ matrix such that $AM_{[a,b]}$ is nonsingular, we know from Section 3.2 that:

$$\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} = -\left(AM_{[a,b]}\right)^{-1} AE\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z) + \frac{\dot{P}(a \leq C \leq b|T_b)}{P(a \leq C \leq b|T_b)} \right\}' \middle| a \leq C \leq b \right].$$

Therefore, the pathwise derivative condition

$$\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} = \left(AM_{[a,b]}\right)^{-1} AE\left[\varphi(O,\beta^0_{[a,b]})S(O)'\right],$$

where $S(O)$ is defined in Section 3.2, will hold if:

$$E\left[\varphi_{[a,b]}(O,\beta^0_{[a,b]})S(O)'\right] = E\left[m \left\{ s(Z) + \frac{\dot{P}(a \leq C \leq b|T_b)}{P(a \leq C \leq b|T_b)} \right\}' \middle| a \leq C \leq b \right],$$

and this is true by Lemma 10. The calculations for this demonstration are tedious and hence they are presented separately under Lemma 10 stated immediately before the present proof.

The pathwise derivative condition holds in the general over identified case ($d_m \geq d_\beta$). Hence, it also holds in the just identified case ($d_m = d_\beta$). To avoid any confusion (at the cost of brevity), we will first complete the proof for case (ii), i.e., the just identified case. We will show that the influence function $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O,\beta^0_{[a,b]})$ obtained from $\varphi_{[a,b]}(O,\beta^0_{[a,b]})$ belongs in $\mathcal{T}$. This follows simply by matching the first set of terms in $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ (i.e., those that correspond to line one in (25)) to the terms corresponding to $\nu_{b+1}(Z_1,\ldots,Z_{b+1}),\ldots,\nu_R(Z_1,\ldots,Z_R)$ in $\mathcal{T}$; the second set of terms (i.e., those that correspond to line two in (25)) to the terms corresponding to $\nu_a(Z_1,\ldots,Z_a),\ldots,\nu_b(Z_1,\ldots,Z_b)$

in $\mathcal{T}$; and the third set of terms (i.e., those that correspond to line three in (25)) to the terms corresponding to $\omega_a(Z_1,\ldots,Z_a),\ldots,\omega_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; while matching zeros with the remaining terms in $\mathcal{T}$. Hence, $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ is the efficient influence function. The expectation of the outer-product of $-M^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ gives the inverse efficiency bound $M^{-1}_{[a,b]}E\left[\varphi_{[a,b]}(O;\beta^0_{[a,b]})\varphi'_{[a,b]}(O;\beta^0_{[a,b]})\right]M^{-1'}_{[a,b]} = M^{-1}_{[a,b]}V_{[a,b]}M^{-1'}_{[a,b]}$.

Now let us get back to the over identified case ($d_m \geq d_\beta$). As noted in Section 3.2, this is where our proof markedly differs from similar proofs in the over identified case because those proofs only do a matching exercise similar to the above without considering the additional restrictions on the tangent set that are imposed by over identification. Arriving at the optimal A, i.e., $M'_{[a,b]}V^{-1}_{[a,b]}$, after this exercise is the same as in Chen et al. (2008) and hence to avoid repetition it is omitted for brevity.

First, consider the case of $a = b$. The above matching also holds with the influence function $-(M'_{[a,a]}V^{-1}_{[a,a]}M_{[a,a]})^{-1}M'_{[a,a]}V^{-1}_{[a,a]}\varphi_{[a,a]}(O,\beta^0_{[a,a]})$. Hence, we focus on verifying the additional restriction (11) due to over identification. If $a = b$ then (11) is:

$$0 = B_{[a,a]}E\left[m(Z;\beta^0_{[a,a]})\left\{\sum_{r=1}^{R}\nu_r(Z_1,\ldots,Z_r) + \omega_a(Z_1,\ldots,Z_a)\right\}'\middle| C = a\right].$$

Therefore, guided exactly by the above matching exercise, $-(M'_{[a,a]}V^{-1}_{[a,a]}M_{[a,a]})^{-1}M'_{[a,a]}V^{-1}_{[a,a]}\varphi_{[a,a]}(O,\beta^0_{[a,a]})$ will satisfy (11) and hence belong in $\mathcal{T}$ if we can show that:

$$\begin{aligned}0 \quad = \quad B_{[a,a]}E\Bigg[m\Bigg\{\sum_{r=1}^{R}\frac{P(C=a|T_a)\left(E[m|T_r]-E[m|T_{r-1}]\right)}{P(C\geq r|T_{r-1})P(C=a)} \\ +\frac{E[m|T_a]}{P(C=a)}\Bigg\}'V^{-1}_{[a,a]}M_{[a,a]}(M'_{[a,a]}V^{-1}_{[a,a]}M_{[a,a]})^{-1}\Bigg| C=a\Bigg].\end{aligned}$$

Now, recalling that $B_{[a,b]} := \left(I_{d_\beta} - M_{[a,b]}\left(AM_{[a,b]}\right)^{-1}A\right)$, it follows that the above equation will hold if:

$$E\left[m\left\{\sum_{r=1}^{R}\frac{P(C=a|T_a)\left(E[m|T_r]-E[m|T_{r-1}]\right)}{P(C\geq r|T_{r-1})P(C=a)} + \frac{E[m|T_a]}{P(C=a)}\right\}'\middle| C=a\right] = V_{[a,a]},$$

which is true since by definition $V_{[a,a]} = E\left[\left\{\varphi_{[a,a]}(O;\beta^0_{[a,a]})\right\}\left\{\varphi_{[a,a]}(O;\beta^0_{[a,a]})\right\}'\right]$, i.e.,

$$\begin{aligned}
V_{[a,a]} &= E\left[\left\{\sum_{r=a+1}^{R}\frac{I(C\geq r)P(C=a|T_a)}{P(C\geq r|T_{r-1})P(C=a)}\left(E[m|T_r]-E[m|T_{r-1}]\right)+\frac{I(C=a)}{P(C=a)}E[m|T_a]\right\}\{\}'\right]\\
&= E\left[\sum_{r=a+1}^{R}\frac{P^2(C=a|T_a)}{P(C\geq r|T_{r-1})P^2(C=a)}m\left(E[m|T_r]-E[m|T_{r-1}]\right)'+\frac{I(C=a)}{P^2(C=a)}mE[m|T_a]'\right]\\
&= E\left[\sum_{r=a+1}^{R}\frac{I(C=a)P(C=a|T_a)}{P(C\geq r|T_{r-1})P^2(C=a)}m\left(E[m|T_r]-E[m|T_{r-1}]\right)'+\frac{I(C=a)}{P^2(C=a)}mE[m|T_a]'\right]\\
&= E\left[m\left\{\sum_{r=1}^{R}\frac{P(C=a|T_a)\left(E[m|T_r]-E[m|T_{r-1}]\right)}{P(C\geq r|T_{r-1})P(C=a)}+\frac{E[m|T_a]}{P(C=a)}\right\}'\middle| C=a\right].
\end{aligned}$$

Now, consider the case of $a=1, b=R$. The matching exercise from the just identified case will not be appropriate here because we have not imposed enough restrictions on the $\omega_r(Z_1,\ldots,Z_r)$'s; see footnote 10. Instead, here we will be guided by the simplified expression of $\varphi_{[1,R]}(O;\beta^0_{[1,R]})$ in Lemma 1, i.e.,

$$\varphi_{[1,R]}(O;\beta^0_{[1,R]}) = \sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r|T_{r-1})}\left(E[m|T_r]-E[m|T_{r-1}]\right)+E[m|T_1],$$

and match the term $-\left(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]}\right)^{-1}M'_{[1,R]}V^{-1}_{[1,R]}\frac{(E[m|T_r]-E[m|T_{r-1}])}{P(C\geq r|T_{r-1})}$ of the influence function $-\left(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]}\right)^{-1}M'_{[1,R]}V^{-1}_{[1,R]}\varphi_{[1,R]}(O;\beta^0_{[1,R]})$ with the term $\nu_r(Z_1,\ldots,Z_r)$ of $\mathcal{T}$ for $r=2,\ldots,R-1$ and the term $-\left(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]}\right)^{-1}M'_{[1,R]}V^{-1}_{[1,R]}E[m|T_1]$ of the influence function $-\left(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]}\right)^{-1}M'_{[1,R]}V^{-1}_{[1,R]}\varphi_{[1,R]}(O;\beta^0_{[1,R]})$ with the term $\nu_1(Z_1)$ of $\mathcal{T}$. Guided exactly by this matching exercise, $-(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]})^{-1}M'_{[1,R]}V^{-1}_{[1,R]}\varphi_{[1,R]}(O,\beta^0_{[1,R]})$ will satisfy (10) and hence belong in $\mathcal{T}$ if we can show that:

$$0 = B_{[1,R]}E\left[m\left\{\sum_{r=1}^{R}\frac{(E[m|T_r]-E[m|T_{r-1}])}{P(C\geq r|T_{r-1})}+E[m|T_1]\right\}'V^{-1}_{[1,R]}M_{[1,R]}(M'_{[1,R]}V^{-1}_{[1,R]}M_{[1,R]})^{-1}\right].$$

Now, recalling that $B_{[1,R]} := \left(I_{d_\beta}-M_{[1,R]}\left(AM_{[1,R]}\right)^{-1}A\right)$, it follows that the above equation will hold if:

$$E\left[m\left\{\sum_{r=1}^{R}\frac{(E[m|T_r]-E[m|T_{r-1}])}{P(C\geq r|T_{r-1})}+E[m|T_a]\right\}'\right]=V_{[1,R]},$$

which it is easy to see is true by following the same steps (but more easily) as done for the case $a=b$. Therefore, we have now established that for both cases $a=b$ and $a=1,b=R$, the influence function $-\Omega_{[a,b]}^{-1}M_{[a,b]}'V_{[a,b]}^{-1}\varphi_{[a,b]}(O;\beta_{[a,b]}^{0})$ belong in the tangent set $\mathcal{T}$. Therefore, $-\Omega_{[a,b]}^{-1}M_{[a,b]}'V_{[a,b]}^{-1}\varphi_{[a,b]}(O;\beta_{[a,b]}^{0})$ is the efficient influence function and hence the expectation of its outer-product gives the inverse efficiency bound $\Omega_{[a,b]}^{-1}$. ■

Proof of Proposition 3: The pathwise derivative condition for this result was verified in Chaudhuri (2020) for the just identified case and applies equally well to the over identified case (similar to what we saw in the proof of Proposition 2). Therefore, we only focus on characterizing the additional restrictions on the tangent set imposed by over identification, and showing that the claimed influence function satisfies those restrictions and thus is the efficient influence function. We hope that it will be clear along the process that the method in Section 3.2 to obtain the additional restrictions, is general enough to obtain the additional restriction on a tangent set that are imposed by over identification in other models too.

Proceeding exactly as in Section 3.2 but, importantly, reflecting the fact that $P(C=r|Z_1,\ldots,Z_r)$ is known, write the log of the distribution of O in terms of $(C,Z')'$ for a regular parametric sub-model indexed by η as (η was θ and Z_r was $Z_{(r)}$ in Chaudhuri (2020)):

$$\log f_\eta(O)=\log f_\eta(Z_1)+\sum_{r=2}^{R}I(C\geq r)\log f_\eta(Z_r|Z_1,\ldots,Z_{r-1})+\sum_{r=1}^{R}I(C=r)\log P(C=r|Z_1,\ldots,Z_r)$$

and the score function with respect to η as:

$$S_\eta(O)=s_\eta(Z_1)+\sum_{r=2}^{R}I(C\geq r)s_\eta(Z_r|Z_1,\ldots,Z_{r-1})$$

where $s_\eta(Z_1):=\frac{\partial}{\partial\eta}\log f_\eta(Z_1)$, and $s_\eta(Z_r|Z_1,\ldots,Z_{r-1}):=\frac{\partial}{\partial\eta}\log f_\eta(Z_r|Z_1,\ \ldots,Z_{r-1})$ for

$r = 2, \ldots, R$. The tangent set is the mean square closure of all d_β dimensional linear combinations of $S_\eta(O)$ for all such smooth parametric sub-models, and it can be generically defined as:

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r)\nu_r(Z_1, \ldots, Z_r), \tag{33}$$

where $\nu_1(Z_1) \in L_0^2(F(Z_1))$ and $\nu_r(Z_1, \ldots, Z_r) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1}))$ for $r = 2, \ldots, R$.

We will proceed as before, but maintain that $P(C = r|Z_1, \ldots, Z_r)$ is known, to obtain the additional restrictions on $\mathcal{T}$ due to over identification. The moment restrictions in (12) give the following identity in η for a given λ:

$$0 \equiv E_\eta[m(Z; \beta_\lambda^0)|C \in \lambda] \equiv E_\eta\left[\frac{P(C \in \lambda|Z)}{P(a \leq C \leq b)} m(Z; \beta_\lambda^0)\right].$$

Differentiate it with respect to η under the integral at $\eta = \eta^0$, and use (1) and (12) to get:

$$0 = M_\lambda \frac{\partial \beta_\lambda^0(\eta_0)}{\partial \eta'} + E\left[m(Z; \beta_\lambda^0)s(Z)'\right| C \in \lambda] \tag{34}$$

where $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$ (with abuse, we briefly revert to the T_r notation for brevity). Now, we note that (12) also gives the following identity in η for given λ:

$$0 \equiv AE_\eta[m(Z; \beta_\lambda^0)|C \in \lambda] \equiv AE_\eta\left[\frac{P(C \in \lambda|Z)}{P(C \in \lambda)} m(Z; \beta_\lambda^0)\right]$$

for any A that is a full row rank $d_\beta \times d_m$ matrix such that AM_λ is nonsingular. Then, as before, solving for $\frac{\partial \beta_\lambda^0(\eta_0)}{\partial \eta'}$, we obtain that:

$$\frac{\partial \beta_\lambda^0(\eta_0)}{\partial \eta'} = -\left(AM_\lambda\right)^{-1} AE\left[m(Z; \beta_\lambda^0)s(Z)'\right| C \in \lambda],$$

which when substituted for in (34) gives (noting that $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$):

$$0 = \left(I_{d_\beta} - M_\lambda \left(AM_\lambda\right)^{-1} A\right) E\left[m(Z; \beta_\lambda^0)\left\{s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})\right\}'\right| C \in \lambda\Bigg].$$

While this is trivially true under just identification, in the case of over identification it implies that the tangent set $\mathcal{T}$ in (33) must satisfy the additional restrictions that

$$0 = \left(I_{d_\beta} - M_\lambda \left(AM_\lambda\right)^{-1} A\right) E\left[m(Z;\beta^0_\lambda) \left\{ \nu(Z_1) + \sum_{r=2}^{R} \nu(Z_1,\ldots,Z_r) \right\}' \middle| C\in\lambda \right]. \qquad (35)$$

Matching terms of $-\bar{\Omega}_\lambda^{-1} M_\lambda' \bar{V}_\lambda^{-1} \bar{\varphi}_\lambda(O;\beta^0_{[a,b]})$ with that of $\mathcal{T}$ where the term involving $\bar{\varphi}_{1,\lambda}$ is matched to $\nu_1(Z_1)$ and the terms involving $\frac{1}{P(C\geq r|T_{r-1})}\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)$ are matched to $\nu_r(Z_1,\ldots,Z_r)$ for $r=2,\ldots,R$, we can say that $-\bar{\Omega}_\lambda^{-1} M_\lambda' \bar{V}_\lambda^{-1} \bar{\varphi}_\lambda(O;\beta^0_{[a,b]}) \in \mathcal{T}$ if additionally

$$0 = \left(I_{d_\beta} - M_\lambda \left(AM_\lambda\right)^{-1} A\right) E\left[m(Z;\beta^0_\lambda) \left\{ \bar{\varphi}_{1,\lambda} + \sum_{r=2}^{R} \frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)}{P(C\geq r|T_{r-1})} \right\}' \bar{V}_\lambda^{-1} M_\lambda \bar{\Omega}_\lambda^{-1} \middle| C\in\lambda \right]$$

which is true since we can easily see that, by repeatedly using (1) and the law of iterated expectations, we can write $\bar{V}_\lambda := E\left[\bar{\varphi}_\lambda(O;\beta^0_{[a,b]})\bar{\varphi}_\lambda(O;\beta^0_{[a,b]})'\right]$ as:

$$\begin{aligned}
\bar{V}_\lambda &= \sum_{r=2}^{R} E\left[\frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)'}{P(C\geq r|T_{r-1})}\right] + E\left[\bar{\varphi}_{1,\lambda}\bar{\varphi}_{1,\lambda}'\right] \\
&= \sum_{r=2}^{R} E\left[\frac{\bar{\varphi}_{r,\lambda}\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)'}{P(C\geq r|T_{r-1})}\right] + E\left[\bar{\varphi}_{1,\lambda}\bar{\varphi}_{1,\lambda}'\right] \\
&= \sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_r)}{P(C\in\lambda)} m \middle| T_r\right]\frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)'}{P(C\geq r|T_{r-1})}\right] + E\left[E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)} m \middle| T_1\right]\bar{\varphi}_{1,\lambda}'\right] \\
&= \sum_{r=2}^{R} E\left[\frac{P(C\in\lambda|T_r)}{P(C\in\lambda)} m \frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)'}{P(C\geq r|T_{r-1})}\right] + E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)} m \bar{\varphi}_{1,\lambda}'\right] \\
&= \sum_{r=2}^{R} E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} m \frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)'}{P(C\geq r|T_{r-1})}\right] + E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} m \bar{\varphi}_{1,\lambda}'\right] \\
&= E\left[m(Z;\beta^0_\lambda) \left\{ \bar{\varphi}_{1,\lambda} + \sum_{r=2}^{R} \frac{\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right)}{P(C\geq r|T_{r-1})} \right\}' \middle| C\in\lambda \right]. \blacksquare
\end{aligned}$$

Proof of Proposition 4: The pathwise derivative condition for this result was verified in Chaudhuri (2020) for the just identified case and applies equally well to the over identified case (similar to that in the proof of Proposition 2). Therefore, we only focus on characterizing the additional restrictions on the tangent set imposed by over identification, and showing that the claimed influence function satisfies those restrictions and thus is the efficient influence function. Proceeding as in Section 3.2 but imposing (13), write the log of the distribution of O in terms of $(C, Z')'$ for a regular parametric sub-model indexed by η as:

$$\log f_\eta(O) = \log f_\eta(Z_1) + \sum_{r=2}^{R} I(C \geq r) \log f_\eta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \log P(C = r|Z_1)$$

and the score function with respect to η as:

$$S_\eta(O) = s_\eta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\eta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\eta(C = r|Z_1)}{P_\eta(C = r|Z_1)}$$

where $s_\eta(Z_1) := \frac{\partial}{\partial\eta} \log f_\eta(Z_1)$, $s_\eta(Z_r|Z_1, \ldots, Z_{r-1}) := \frac{\partial}{\partial\eta} \log f_\eta(Z_r|Z_1, \ \ldots, Z_{r-1})$ for $r = 2, \ldots, R$, and $\dot{P}_\eta(C = r|Z_1) := \frac{\partial}{\partial\eta} P_\eta(C = r|Z_1)$ for $r = 1, \ldots, R$. We will actually use an apparently cumbersome but ultimately more convenient representation of the score function by using the two equivalent factorizations of the joint distribution of $I(C \in \lambda)$ and Z_1:

$$\begin{aligned} & s_\eta(Z_1) + I(C \in \lambda) \frac{\dot{P}_\eta(C \in \lambda|Z_1)}{P_\eta(C \in \lambda|Z_1)} + I(C \notin \lambda) \frac{\dot{P}_\eta(C \notin \lambda|Z_1)}{P_\eta(C \notin \lambda|Z_1)} \\ &= I(C \in \lambda) \left[\frac{\dot{P}_\eta(C \in \lambda)}{P_\eta(C \in \lambda)} + s_\eta(Z_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}_\eta(C \notin \lambda)}{P_\eta(C \notin \lambda)} + s_\eta(Z_1|C \notin \lambda) \right] \end{aligned} \quad (36)$$

where $s_\eta(Z_1|C \in \lambda) := \frac{\partial}{\partial\eta} \log f_\eta(Z_1|C \in \lambda)$, $s_\eta(Z_1|C \notin \lambda) := \frac{\partial}{\partial\eta} \log f_\eta(Z_1|C \notin \lambda)$, $\dot{P}_\eta(C \in \lambda|Z_1) := \frac{\partial}{\partial\eta} P_\eta(C \in \lambda|Z_1) =: -\dot{P}_\eta(C \notin \lambda|Z_1)$ and $\dot{P}_\eta(C \in \lambda) := \frac{\partial}{\partial\eta} P_\eta(C \in \lambda) =: -\dot{P}_\eta(C \notin \lambda)$. Then substituting for $s_\eta(Z_1)$ in $S_\eta(O)$ we obtain the cumbersome but useful expression:

$$\begin{aligned} S_\eta(O) \quad = \quad & I(C \in \lambda) \left[\frac{\dot{P}_\eta(C \in \lambda)}{P_\eta(C \in \lambda)} + s_\eta(Z_1|C \in \lambda) - \frac{\dot{P}_\eta(C \in \lambda|Z_1)}{P_\eta(C \in \lambda|Z_1)} \right] \\ & + I(C \notin \lambda) \left[\frac{\dot{P}_\eta(C \in \lambda)}{P_\eta(C \in \lambda) - 1} + s_\eta(Z_1|C \notin \lambda) - \frac{\dot{P}_\eta(C \in \lambda|Z_1)}{P_\eta(C \in \lambda|Z_1) - 1} \right] \\ & + \sum_{r=2}^{R} I(C \geq r) s_\eta(Z_r|Z_1, \ldots, Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\eta(C = r|Z_1)}{P_\eta(C = r|Z_1)}. \end{aligned}$$

Hence the representation of the tangent set that we will consider is:

$$
\begin{aligned}
\mathcal{T} \; &:= \; I(C \in \lambda)\left[\frac{a}{b} + \mu_1(Z_1, C \in \lambda) - \frac{a(Z_1)}{b(Z_1)}\right] + I(C \notin \lambda)\left[\frac{a}{b-1} + \mu_2(Z_1, C \notin \lambda) - \frac{a(Z_1)}{b(Z_1)-1}\right] \\
&\quad + \sum_{r=2}^{R} I(C \geq r)\nu_r(Z_1, \ldots, Z_r) + \sum_{r=1}^{R} I(C = r)\omega_r(Z_1), \qquad (37)
\end{aligned}
$$

where a and $b \in (0,1)$ are constants; $a(z_1)$ and $b(Z_1)$ are such that $a(Z_1)/b(Z_1)$ and $a(Z_1)/(b(Z_1)-1)$ are square intergrable functions of Z_1; $\mu_1(Z_1, C \in \lambda) \in L_0^2(F(Z_1|C \in \lambda))$ and $\mu_2(Z_1, C \notin \lambda) \in L_0^2(F(Z_1|C \notin \lambda))$; and the terms described so far satisfy the restriction that the first line on the RHS of (37) is $L_0^2(F(Z_1))$ (since it represents $s(Z_1)$); whereas $\nu_r(Z_1, \ldots, Z_r) \in L_0^2(F(Z_r|Z_1, \ldots, Z_{r-1}))$ for $r = 2, \ldots, R$, and $\omega_r(Z_1)$ is any square integrable function of Z_1 for $r = 1, \ldots, R$. To obtain the additional restrictions due to over identification of β_λ^0, we write $\left(I_{d_\beta} - M_\lambda \left(AM_\lambda\right)^{-1} A\right)$ as B_λ for brevity, and then imposing CMAR (13) we arrive at the counterpart of (9) for a given λ as:

$$
0 \;=\; B_\lambda E\left[m(Z;\beta_\lambda^0)\left\{ s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1}) + \frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)}\right\}' \middle| C \in \lambda \right]
$$

which gives the additional restrictions on $\mathcal{T}$ in (37) as:

$$
0 = B_\lambda\left\{ E\left[m(Z;\beta_\lambda^0)\sum_{r=2}^{R}\nu_r(Z_1, \ldots, Z_r)' \middle| C \in \lambda\right] + E\left[m(Z;\beta_\lambda^0)\frac{I(C \in \lambda)}{P(C \in \lambda)}\left\{ s(Z_1) + \frac{\dot{P}(C \in \lambda|Z_1)}{P(C \in \lambda|Z_1)}\right\}'\right]\right\}.
$$

Substitute for $I(C \in \lambda)\left\{s(Z_1) + \dot{P}(C \in \lambda|Z_1)/P(C \in \lambda|Z_1)\right\}$ from (36) to get:

$$
\begin{aligned}
0 \;&=\; B_\lambda E\left[m\sum_{r=2}^{R}\nu_r(Z_1, \ldots, Z_r)' \middle| C \in \lambda\right] + B_\lambda E\left[m\frac{I(C \in \lambda)}{P(C \in \lambda)}\left\{ s(Z_1|C \in \lambda) + \frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)}\right\}'\right] \\
&=\; B_\lambda E\left[m\sum_{r=2}^{R}\nu_r(Z_1, \ldots, Z_r)' \middle| C \in \lambda\right] + B_\lambda E\left[m\frac{I(C \in \lambda)}{P(C \in \lambda)} s(Z_1|C \in \lambda)'\right] \qquad (38)
\end{aligned}
$$

using the moment restrictions in (12). (We are writing $m(Z;\beta_\lambda^0)$ as m for brevity.) Hence, over identification of β^0_λ imposes the additional restrictions (38) on $\mathcal{T}$ in (37).

Now, match the terms of $-\left[\Omega^{CMAR}_\lambda\right]^{-1} M'_\lambda \left[V^{CMAR}_\lambda\right]^{-1} \varphi^{CMAR}_\lambda(O;\beta^0_{[a,b]})$ with the terms of $\mathcal{T}$ as follows. The terms involving $\frac{P(C\in\lambda|T_1)}{P(C\geq r|T_{r-1})P(C\in\lambda)}\left(E[m|T_r]-E[m|T_{r-1}]\right)$ are matched to $\nu_r(Z_1,\ldots,Z_r)$ for $r=2,\ldots,R$. The term involving $\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]$ is matched to $I(C\in\lambda)s(Z_{(1)}|C\in\lambda)$. The other terms in $\mathcal{T}$ are matched to zeros. Therefore, the influence function $\left[\Omega^{CMAR}_\lambda\right]^{-1} M'_\lambda \left[V^{CMAR}_\lambda\right]^{-1} \varphi^{CMAR}_\lambda(O;\beta^0_{[a,b]})$ will belong in $\mathcal{T}$ and hence will be the efficient influence function if additionally:

$$\begin{aligned}0 \;=\; B_\lambda\Bigg\{&E\left[m\sum_{r=2}^{R}\frac{P(C\in\lambda|T_1)}{P(C\geq r|T_{r-1})P(C\in\lambda)}\left(E[m|T_r]-E[m|T_{r-1}]\right)'\middle| C\in\lambda\right]\\ &+E\left[m\frac{I(C\in\lambda)}{P^2(C\in\lambda)}E[m|T_1]'\right]\Bigg\}\left[V^{CMAR}_\lambda\right]^{-1}M_\lambda\left[\Omega^{CMAR}_\lambda\right]^{-1}\end{aligned}$$

i.e., if:

$$V^{CMAR}_\lambda = E\left[m\sum_{r=2}^{R}\frac{P^2(C\in\lambda|T_1)}{P(C\geq r|T_{r-1})P^2(C\in\lambda)}\left(E[m|T_r]-E[m|T_{r-1}]\right)'\right]+E\left[m\frac{I(C\in\lambda)}{P^2(C\in\lambda)}E[m|T_1]'\right],$$

which it can be seen is true by writing out the expression for $V^{CMAR}_\lambda := Var(\varphi^{CMAR}_\lambda(O;\beta^0_\lambda))$ and then using CMAR in (13) and the law of iterated expectations as in the last proof. ■

Proof of Lemma 5: Note that:

$$\begin{aligned}\omega^{\text{IPW}}_{[a,b]} &:= \frac{I(C=R)}{\prod_{r=1}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{\sum_{j=a}^{b}P(C=j|T_j,C\geq j)\prod_{k=1}^{j-1}(1-P(C=k|T_k,C\geq k))}{P(a\leq C\leq b)}\\ &= \sum_{j=a}^{b}\frac{I(C=R)}{\prod_{r=1}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{P(C=j|T_j,C\geq j)\prod_{k=1}^{j-1}(1-P(C=k|T_k,C\geq k))}{P(a\leq C\leq b)}\\ &= \sum_{j=a}^{b}\frac{I(C=R)}{P(C=R|T_R)}\frac{P(C=j|T_j)}{P(a\leq C\leq b)}\qquad\left[=\sum_{j=a}^{b}\frac{P(C=j)}{P(a\leq C\leq b)}\frac{I(C=R)}{P(C=R|T_R)}\frac{P(C=j|T_j)}{P(C=j)}\right]\\ &= \frac{I(C=R)}{P(C=R|T_R)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)} = \frac{I(C=R)}{P(C=R|T_R)}\frac{P(a\leq C\leq b|T_R)}{P(a\leq C\leq b)}\end{aligned}$$

where the last two equalities follow by (1). Therefore, since $Z\equiv T_R$, it follows by using the law of iterated expectations in the second and third equalities below, that:

$$
\begin{aligned}
E\left[\omega^{\text{IPW}}_{[a,b]} m(Z;\beta)\right] &= E\left[\frac{I(C=R)}{P(C=R|T_R)} \frac{P(a \le C \le b|T_R)}{P(a \le C \le b)} m(T_R;\beta)\right] \\
&= E\left[\frac{P(a \le C \le b|T_R)}{P(a \le C \le b)} m(T_R;\beta)\right] \\
&= E\left[\frac{I(a \le C \le b)}{P(a \le C \le b)} m(T_R;\beta)\right] \\
&= E[m(Z;\beta)|a \le C \le b]. \ \blacksquare
\end{aligned}
$$

Proof of Proposition 6: (i) will follow if Condition 1 of Ackerberg et al. (2014) holds. Our assumptions A1 and A3 directly imply Condition 1(i) and 1(ii) hold. Furthermore, Condition 1(iii) also holds by virtue of our assumption A2 because for any $r = a, \ldots, R-1$:

$$\frac{\partial}{\partial p_r(T_r)} E\left[I(C \ge r)\left\{I(C=r) - p_r(T_r)\right\}|T_r\right] = -P(C \ge r|T_r) \ne 0 \quad \text{a.s. } T_r.$$

Before proceeding further, we note using the expression in (5) and Lemma 5 that:

$$I(C=R)\omega^{IPW}_{[a,b]} m(Z;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} I(C=R)\omega^{IPW}_{[j,j]} m(Z;\beta) \tag{39}$$

and hence for the sake of a cleaner proof it is useful to work on:

$$I(C=R)\omega^{IPW}_{[j,j]} m(Z;\beta) \quad \text{where} \quad \omega^{IPW}_{[j,j]} = \frac{P(C=j|T_j, C \ge j)}{P(C=j)\prod_{k=j}^{R-1}[1 - P(C=k|T_k, C \ge k)]}$$

and then combine the results based on the weights $P(C=j)/P(a \le C \le b)$.

For any $j = a, \ldots, b$ replace $P(C = r|T_r, C \ge r)$ by $h_{j,r}(T_r) := 1/(1 - p_r(T_r))$ for $r = j+1, \ldots, R-1$ and $P(C = j|T_j, C \ge j)$ by $h_{j,j}(T_j) := p_j(T_j)/(1 - p_j(T_j))$ in $\omega^{IPW}_{[j,j]}$ to define (the reason behind the double subscript j, r in h will be clear soon):

$$\phi_{[j,j]}(C, T_R; \beta, h_{j,j}(T_j), \ldots, h_{j,R-1}(T_{R-1})) := I(C=R)\frac{\prod_{k=j}^{R-1} h_{j,k}(T_j)}{P(C=j)} m(T_R;\beta). \tag{40}$$

Let $h^0_{j,r}(T_r) := 1/(1 - P(C = r|T_r, C \geq r))$ for $r = j+1, \ldots, R-1$ and $h^0_{j,j}(T_j) = P(C = j|T_j, C \geq j)/(1 - P(C = j|T_j, C \geq j))$. Then, trivially $\frac{\partial E[\phi_{[j,j]}(C,T_R;\beta,h^0_{j,j}(T_j),\ldots,h^0_{j,R-1}(T_{R-1}))]}{\partial h_{j,r}}[.]$ is a linear functional for $r = j, \ldots, R-1$. We maintain the assumption that it is also a bounded functional as defined in Ackerberg et al. (2014). (The boundedness is maintained as a high level assumptions since under our assumption A2 it can hold in various ways depending on the interplay between the $E[m|T_r]$'s and the conditional hazards; e.g., taking $j = 1, R = 2$, we can see that $\frac{\partial E[\phi_{[1,1]}(C,T_2;\beta,h_{1,1})]}{\partial h_{1,1}} = E[P(C = 2|T_1)m(Z; \beta^0_{[1,R]})/P(C = 1)]$.) Thus, Condition 1(iv) of Ackerberg et al. (2014) also holds under our maintained assumptions. However, our interest is not always on a unitary sub-population $[j, j]$ but more generally on $[a, b]$, and for that we know from (39) that we should be looking at:

$$\sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \phi_{[j,j]}(C, T_R; \beta, h_{j,j}(T_j), \ldots, h_{j,R-1}(T_{R-1})).$$

Before proceeding further we remark here about the double subscript in h. We redefined the nuisance parameters as h to make the functionals linear in h. However, the h's that enter the above linear combination are not unique — $h_{a,k}(T_k) = \ldots = h_{k-1,k}(T_k)$ for any $k = a+1, \ldots, b$ and $h_{a,k}(T_k) = \ldots = h_{k-1,k}(T_k)$ for any $k = b+1, \ldots, R-1$, while $h_{j,k}(T_k)$ appearing in $\phi_{[j,j]}(.)$ and $h_{k,k}(T_k)$ appearing in $\phi_{[k,k]}(.)$ for $k = j+1, \ldots, b$ and $j = a \ldots, b-1$ both depend on $p_k(T_k)$ only but in different ways. Pretending that the h's are distinct does not cause any problem, not even with the invertibility in Condition 1(iii) of Ackerberg et al. (2014) since that will lead to a diagonal matrix (and it will be important to keep this last statement in mind for the proof of part (ii)). Therefore if Condition 1 of Ackerberg et al. (2014) holds for $\phi_{[j,j]}(.)$ for $j = a, \ldots, b$, which we have already shown, then it also holds for the above linear combination those $\phi_{[j,j]}(.)$'s. This completes the proof of part (i).

(ii) It is straightforward to see that:

$$E\left[\frac{\partial}{\partial \beta'} \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \phi_{[j,j]}(C, T_R; \beta^0_{[a,b]}, h_{j,j}(T_j), \ldots, h_{j,R-1}(T_{R-1}))\right] = M_{[a,b]}.$$

Hence, we know from Theorem 1 of Ackerberg et al. (2014) that the efficiency bound for $\beta^0_{[a,b]}$ based on the information contained in *only* the moment restrictions in part (i) is:

$$\widetilde{\Omega}_{[a,b]} := M'_{[a,b]}\left[Var\left(\sum_{j=a}^{b}\frac{P(C=j)}{P(a\le C\le b)}\widetilde{\phi}_{[j,j]}(C,T_R;\beta^0_{[a,b]},h^0_{j,j}(T_j),\dots,h^0_{j,R-1}(T_{R-1}))\right)\right]^{-1}M_{[a,b]} \tag{41}$$

where, writing $(h_{j,j}(T_j),\dots,h_{j,R-1}(T_{R-1}))$ as $h_{j,j:R-1}(T_{R-1})$ and its true value as $h^0_{j,j:R-1}(T_{R-1})$:

$$\widetilde{\phi}_{[j,j]}(C,T_R;\beta,h_{j,j:R-1}(T_{R-1})) := \phi_{[j,j]}(C,T_R;\beta,h_{j,j:R-1}(T_{R-1}))-\sum_{k=j}^{R-1}\frac{D^0_{j,k}(T_k;\beta)}{S^0_{j,k}(T_k)}s_{j,k}(C,T_k,h_{j,k}(T_k)) \tag{42}$$

and where $D^0_{j,k}(T_k;\beta)$, $S^0_{j,k}(T_k)$ and $s_{j,k}(C,T_k,h_{j,k}(T_k))$ are as follows. For $j=a,\dots,b$:

$$\begin{aligned} s_{j,k}(C,T_k,h_{j,k}(T_k)) &:= I(C\ge k)\left[I(C=k)-\frac{h_{j,k}(T_k)-1}{h_{j,k}(T_k)}\right] && \text{for } k=j+1,\dots,R-1\\ &:= I(C\ge k)\left[I(C=k)-\frac{h_{j,k}(T_k)}{1+h_{j,k}(T_k)}\right] && \text{for } k=j \end{aligned}$$

whereas for $j=a,\dots,b$ and $k=j,\dots,R-1$:

$$S^0_{j,k}(T_k) := \frac{\partial E[s_{j,k}(C,T_k,h^0_{j,k}(T_k))]}{\partial h_{j,k}} = -P(C\ge|k|T_k)\left(1-P(C=k|T_k,C\ge k)\right)^2.$$

$D^0_{j,k}(T_k;\beta)v_{j,k}(T_k)$ is the pathwise derivative of $E[\phi_{[j,j]}(C,T_R;\beta,h_{j,j:R-1}(T_{R-1}))|T_k]$ with respect to $h_{j,k}(T_k)$ in the direction $v_{j,k}(T_k)\in H_{j,k}(T_k)-\{h^0_{j,k}(T_k)\}$ (where $\mathcal{H}_{j,k}(T_k)$ is the function space for $h_{j,k}(T_k)$) evaluated at $h^0_{j,j:R-1}(T_{R-1})$, i.e., for $j=a,\dots,b$ and $k=j,\dots,R-1$:

$$D^0_{j,k}(T_k;\beta)v_{j,k}(T_k) = \frac{\partial E[\phi_{[j,j]}(C,T_R;\beta,h^0_{j,j:R-1}(T_{R-1}))|T_k]}{\partial h_{j,k}}[v_{j,k}].$$

First, note that:

$$D^0_{j,k}(T_k;\beta) = E\left[\left\{\prod_{r=j,\dots,R-1;r\neq k}h_{j,r}(T_r)\right\}I(C=R)\frac{m(Z;\beta)}{P(C=j)}\middle|T_k\right]$$

i.e., for $k = j+1, \dots, R-1$:

$$
\begin{aligned}
D^0_{j,k}(T_k;\beta) &= E\left[\left\{\prod_{r=j}^{R-1} h_{j,r}(T_r)\right\} I(C=R)\frac{m(Z;\beta)}{P(C=j)h_{j,k}(T_k)}\middle| T_k\right] \\
&= E\left[\frac{I(C=R)P(C=j|T_j,C\geq j)}{\prod_{r=j}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{m(Z;\beta)}{P(C=j)h_{j,k}(T_k)}\middle| T_k\right] \\
&= E\left[\frac{I(C=R)P(C=j|T_j,C\geq j)\prod_{r=1}^{j-1}(1-P(C=r|T_r,C\geq r))}{\prod_{r=1}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{m(Z;\beta)}{P(C=j)h_{j,k}(T_k)}\middle| T_k\right] \\
&= E\left[\frac{I(C=R)P(C=j|T_j)}{P(C=R|T_{R-1})}\frac{m(Z;\beta)}{P(C=j)h_{j,k}(T_k)}\middle| T_k\right] \\
&= E\left[\frac{m(Z;\beta)P(C=j|T_j)}{P(C=j)h_{j,k}(T_k)}\middle| T_k\right] \\
&= \frac{E[m(Z;\beta)|T_k]P(C=j|T_j)(1-P(C=k|T_k,C\geq k))}{P(C=j)}
\end{aligned}
$$

where the second last equality follows by (1) and the law of iterated expectations, whereas:

$$
\begin{aligned}
D^0_{j,j}(T_j;\beta) &= E\left[\left\{\prod_{r=j+1}^{R-1} h_{j,r}(T_r)\right\} I(C=R)\frac{m(Z;\beta)}{P(C=j)}\middle| T_j\right] \\
&= E\left[\frac{I(C=R)}{\prod_{r=j+1}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{m(Z;\beta)}{P(C=j)}\middle| T_j\right] \\
&= E\left[\frac{I(C=R)\prod_{r=1}^{j}(1-P(C=r|T_r,C\geq r))}{\prod_{r=1}^{R-1}(1-P(C=r|T_r,C\geq r))}\frac{m(Z;\beta)}{P(C=j)}\middle| T_j\right] \\
&= E\left[\frac{I(C=R)P(C\geq j+1|T_j)}{P(C=R|T_{R-1})}\frac{m(Z;\beta)}{P(C=j)}\middle| T_j\right] \\
&= E\left[P(C\geq j+1|T_j)\frac{m(Z;\beta)}{P(C=j)}\middle| T_j\right] \\
&= \frac{E[m(Z;\beta)|T_j]P(C\geq j+1|T_j)}{P(C=j)}
\end{aligned}
$$

where, as before, the second last equality follows by (1) and the law of iterated expectations.

Plugging them in (42) at $\beta^0_{[a,b]}, h^0_{j,j:R-1}(T_{R-1})$ gives:

$$
\widetilde{\phi}_{[j,j]}(C,T_R;\beta^0_{[a,b]},h^0_{j,j:R-1}(T_{R-1})) = \phi_{[j,j]}(C,T_R;\beta^0_{[a,b]},h^0_{j,j:R-1}(T_{R-1})) - \sum_{k=j}^{R-1}\frac{D^0_{j,k}(T_k;\beta)}{S^0_{j,k}(T_k)}s_{j,k}(C,T_k,h^0_{j,k}(T_k))
$$

where, writing $m(Z;\beta^0_{[a,b]})$ as m for brevity, the RHS of the above equation is:

$$\begin{aligned}
& I(C=R)\omega^{IPW}_{[j,j]}m - \sum_{k=j+1}^{R-1} \frac{D^0_{j,k}(T_k;\beta)}{S^0_{j,k}(T_k)} s_{j,k}(C,T_k,h^0_{j,k}(T_k)) - \frac{D^0_{j,j}(T_k;\beta)}{S^0_{j,j}(T_k)} s_{j,j}(C,T_k,h^0_{j,j}(T_j)) \\
= \; & \varphi_{[j,j]}(O;\beta^0_{[a,b]}) \qquad (43)
\end{aligned}$$

by (6) because we know from the above calculations that for $k = j+1,\ldots,R-1$:

$$\begin{aligned}
& -\frac{D^0_{j,k}(T_k;\beta)}{S^0_{j,k}(T_k)} s_{j,k}(C,T_k,h^0_{j,k}(T_k)) \\
= \; & \frac{E[m|T_k]\frac{P(C=j|T_j)}{P(C=j)}(1-P(C=k|T_k,C\geq k))}{P(C\geq k|T_k)(1-P(C=k|T_k,C\geq k))^2}\left[I(C=k)-I(C\geq k)P(C=k|T_k,C\geq k)\right] \\
= \; & \frac{E[m|T_k]\frac{P(C=j|T_j)}{P(C=j)}}{P(C\geq k|T_k)(1-P(C=k|T_k,C\geq k))}\left[I(C\geq k)-I(C\geq k+1)-I(C\geq k)P(C=k|T_k,C\geq k)\right] \\
= \; & \left[\frac{I(C\geq k)}{P(C\geq k|T_k)} - \frac{I(C\geq k+1)}{P(C\geq k|T_k)(1-P(C=k|T_k,C\geq k))}\right]\frac{P(C=j|T_j)}{P(C=j)}E[m|T_j] \\
= \; & \left[\frac{I(C\geq k)}{P(C\geq k|T_k)} - \frac{I(C\geq k+1)}{P(C\geq k+1|T_k)}\right]\frac{P(C=j|T_j)}{P(C=j)}E[m|T_j] \\
= \; & \left[\frac{I(C\geq k)}{P(C\geq k|T_{k-1})} - \frac{I(C\geq k+1)}{P(C\geq k+1|T_k)}\right]\frac{P(C=j|T_j)}{P(C=j)}E[m|T_j] \quad \text{[by Lemma 8]}
\end{aligned}$$

whereas for $k=j$:

$$\begin{aligned}
& -\frac{D^0_{j,j}(T_j;\beta)}{S^0_{j,j}(T_j)} s_{j,j}(C,T_j,h^0_{j,j}(T_j)) \\
= & \frac{E[m|T_j]\frac{P(C\geq j+1|T_j)}{P(C=j)}}{P(C\geq j|T_j)(1-P(C=j|T_j,C\geq j))^2}\left[I(C=j)-I(C\geq j)P(C=j|T_j,C\geq j)\right] \\
= & \frac{E[m|T_j]\frac{P(C\geq j+1|T_j)}{P(C=j)}}{P(C\geq j|T_j)(1-P(C=j|T_j,C\geq j))^2}\left[I(C=j)-\{I(C=j)+I(C\geq j+1)\}P(C=j|T_j,C\geq j)\right] \\
= & \left[\frac{I(C=j)P(C\geq j+1|T_j)}{P(C\geq j|T_j)(1-P(C=j|T_j,C\geq j))} - \frac{I(C\geq j+1)P(C\geq j+1|T_j)P(C=j|T_j,C\geq j)}{P(C\geq j|T_j)(1-P(C=j|T_j,C\geq j))^2}\right]\frac{E[m|T_j]}{P(C=j)} \\
= & \left[\frac{I(C=j)P(C\geq j+1|T_j)}{P(C\geq j+1|T_j)} - \frac{I(C\geq j+1)P(C\geq j+1|T_j)\frac{P(C=j|T_j)}{P(C\geq j|T_j)}}{P(C\geq j+1|T_j)\frac{P(C\geq j|T_j)-P(C=j|T_j)}{P(C\geq j|T_j)}}\right]\frac{E[m|T_j]}{P(C=j)} \\
= & \left[I(C=j)-I(C\geq j+1)\frac{P(C=j|T_j)}{P(C\geq j+1|T_j)}\right]\frac{E[m|T_j]}{P(C=j)}.
\end{aligned}$$

Therefore, using (43) and (6) for the first equality and then (3) for the second, imply that:

$$\sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \widetilde{\phi}_{[j,j]}(C, T_R; \beta^0_{[a,b]}, h^0_{j,j:R-1}(T_{R-1})) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \varphi_{[j,j]}(O; \beta^0_{[a,b]}) = \varphi_{[a,b]}(O; \beta^0_{[a,b]}).$$

Hence, by (41) we obtain that $\tilde{\Omega}_{[a,b]} = \Omega_{[a,b]}$ as defined in Proposition 2. ■

Proof of Lemma 7: (i) The equivalence of the limited and full information approach here follows exactly as in part (i) of Proposition 6, with the only change that the conditional hazards are all now conditioned on T_1 only. Consequently, the influence function in part (i) will take a different form here, and for the rest of the proof part (i) we derive that form. To avoid introducing new notation we follow the notation from the last proposition as much as we can. We know from Theorem 1 of Ackerberg et al. (2014) that the efficiency bound for $\beta^0_{[a,b]}$ based on the information contained *only* in the moment restrictions in part (i) is:

$$M'_{[a,b]} \left[Var \left(\sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \widetilde{\phi}_{[j,j]}(C, T_R; \beta^0_{[a,b]}, h^0_{j,j}(T_1), \ldots, h^0_{j,R-1}(T_1)) \right) \right]^{-1} M_{[a,b]}$$

where, writing $(h_{j,j}(T_1), \ldots, h_{j,R-1}(T_1))$ as $h_{j,j:R-1}(T_1)$ and its true value as $h^0_{j,j:R-1}(T_1)$:

$$\begin{aligned}
\widetilde{\phi}_{[j,j]}(C, T_R; \beta, h_{j,j:R-1}(T_1)) \quad &:= \quad \phi_{[j,j]}(C, T_R; \beta, h_{j,j:R-1}(T_1)) - \sum_{k=j}^{R-1} \frac{D^0_{j,k}(T_1; \beta)}{S^0_{j,k}(T_1)} s_{j,k}(C, T_1, h_{j,k}(T_1)) \\
\phi_{[j,j]}(C, T_R; \beta, h_{j,j:R-1}(T_1)) \quad &:= \quad I(C=R) \frac{\prod_{k=j}^{R-1} h_{j,k}(T_1)}{P(C=j)} m(T_R; \beta) \\
s_{j,k}(C, T_1, h_{j,k}(T_1)) \quad &:= \quad I(C \ge k) \left[I(C=k) - \frac{h_{j,k}(T_1) - 1}{h_{j,k}(T_1)} \right] \quad \text{for } k = j+1, \ldots, R-1 \\
&:= \quad I(C \ge k) \left[I(C=k) - \frac{h_{j,k}(T_1)}{1 + h_{j,k}(T_1)} \right] \quad \text{for } k = j \\
S^0_{j,k}(T_1) \quad &:= \quad \frac{\partial E[s_{j,k}(C, T_1, h^0_{j,k}(T_1))]}{\partial h_{j,1}} = -P(C \ge |k|T_1) \left(1 - P(C=k|T_1, C \ge k)\right)^2
\end{aligned}$$

for $j = a, \ldots, b$ and $k = j, \ldots, R-1$. $D^0_{j,k}(T_1; \beta) v_{j,k}(T_1)$ is the pathwise derivative of $E[\phi_{[j,j]}(C, T_R; \beta, h_{j,j:R-1}(T_1))|T_1]$ with respect to $h_{j,k}(T_1)$ in the direction $v_{j,k}(T_1) \in H_{j,k}(T_1)-$

$\{h^0_{j,k}(T_1)\}$ (where $\mathcal{H}_{j,k}(T_1)$ is the function space for $h_{j,k}(T_1)$) evaluated at $h^0_{j,j:R-1}(T_1)$, i.e.,

$$D^0_{j,k}(T_1;\beta)v_{j,k}(T_1) = \frac{\partial E[\phi_{[j,j]}(C,T_R;\beta,h^0_{j,j:R-1}(T_1))|T_1]}{\partial h_{j,k}}[v_{j,k}] \quad \text{for } j=a,\dots,b \text{ and } k=j,\dots,R-1.$$

Therefore, just like before (but now with conditioning set T_1 for all terms):

$$\begin{aligned} D^0_{j,k}(T_1;\beta) &= \frac{E[m(Z;\beta)|T_1]P(C=j|T_1)(1-P(C=k|T_1,C\geq k))}{P(C=j)} \quad \text{for } k=j+1,\dots,R-1, \\ D^0_{j,j}(T_1;\beta) &= \frac{E[m(Z;\beta)|T_1]P(C\geq j+1|T_1)}{P(C=j)}, \end{aligned}$$

and hence for $k=j+1,\dots,R-1$:

$$-\frac{D^0_{j,k}(T_1;\beta)}{S^0_{j,k}(T_1)}s_{j,k}(C,T_1,h^0_{j,k}(T_1)) = \left[\frac{I(C\geq k)}{P(C\geq k|T_1)} - \frac{I(C\geq k+1)}{P(C\geq k+1|T_1)}\right]\frac{P(C=j|T_1)}{P(C=j)}E[m|T_1]$$

whereas for $k=j$:

$$-\frac{D^0_{j,j}(T_1;\beta)}{S^0_{j,j}(T_1)}s_{j,j}(C,T_1,h^0_{j,j}(T_1)) = \left[I(C=j) - I(C\geq j+1)\frac{P(C=j|T_1)}{P(C\geq j+1|T_1)}\right]\frac{E[m|T_1]}{P(C=j)},$$

and therefore:

$$-\sum_{k=j}^{R-1}\frac{D^0_{j,k}(T_1;\beta)}{S^0_{j,k}(T_1)}s_{j,k}(C,T_1,h^0_{j,k}(T_1)) = \left\{\frac{I(C=j)}{P(C=j)} - \frac{I(C=R)}{P(C=R|T_1)}\frac{P(C=j|T_1)}{P(C=j)}\right\}E[m|T_1]$$

which gives:

$$\begin{aligned} &\widetilde{\phi}_{[j,j]}(C,T_R;\beta^0_{[a,b]},h^0_{j,j:R-1}(T_1)) \\ &= \frac{I(C=R)}{P(C=R|T_1)}\frac{P(C=j|T_1)}{P(C=j)}m + \left\{\frac{I(C=j)}{P(C=j)} - \frac{I(C=R)}{P(C=R|T_1)}\frac{P(C=j|T_1)}{P(C=j)}\right\}E[m|T_1]. \end{aligned}$$

Therefore,

$$\sum_{j=a}^{n}\frac{P(C=j)}{P(a\leq C\leq b)}\widetilde{\phi}_{[j,j]}(C,T_R;\beta^0_{[a,b]},h^0_{j,j:R-1}(T_1)) = \varphi^\dagger_{[a,b]}.$$

Adding and subtracting the same terms to $\varphi^{\dagger}_{[a,b]}$ in order to match $\varphi^{CMAR}(O;\beta^{0}_{[a,b]})$ from Proposition 4, we obtain:

$$\begin{aligned}
\varphi^{\dagger}_{[a,b]} &= \sum_{r=2}^{R} \frac{I(C \geq R)}{P(C \geq R|T_1)} \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} \left(E[m|T_r] - E[m|T_{r-1}]\right) \\
&\quad + \left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} - \left(\frac{I(C \geq R)}{P(C \geq R|T_1)} - \frac{I(C \geq R)}{P(C \geq R|T_1)}\right) \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)}\right] E[m|T_1] \\
&= \sum_{r=2}^{R} \frac{I(C \geq R)}{P(C \geq R|T_1)} \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} \left(E[m|T_r] - E[m|T_{r-1}]\right) + \frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} E[m|T_1].
\end{aligned}$$

(ii) Taking variance, we obtain:

$$V^{\dagger}_{[a,b]} = \sum_{r=2}^{R} E\left[\frac{P^2(a \leq C \leq b|T_1)}{P(C \geq R|T_1) P^2(a \leq C \leq b)} Var\left(E[m|T_r]|T_{r-1}\right)\right] + E\left[\frac{I(a \leq C \leq b)}{P^2(a \leq C \leq b)} E[m|T_1] E'[m|T_1]\right]$$

whereas we know from Proposition 4 that:

$$V^{CMAR}_{[a,b]} = \sum_{r=2}^{R} E\left[\frac{P^2(a \leq C \leq b|T_1)}{P(C \geq r|T_1) P^2(a \leq C \leq b)} Var\left(E[m|T_r]|T_{r-1}\right)\right] + E\left[\frac{I(a \leq C \leq b)}{P^2(a \leq C \leq b)} E[m|T_1] E'[m|T_1]\right]$$

Therefore, we obtain that $V^{\dagger}_{[a,b]} - V^{CMAR}_{[a,b]}$ is:

$$\begin{aligned}
& \sum_{r=2}^{R} E\left[\frac{P^2(a \leq C \leq b|T_1)}{P^2(a \leq C \leq b)} \left[\frac{1}{P(C \geq R|T_1)} - \frac{1}{P(C \geq r|T_1)}\right] Var\left(E[m|T_r]|T_{r-1}\right)\right] \\
= \ & \sum_{r=2}^{R} E\left[\frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} \left[\frac{1}{P(C \geq R|T_1)} - \frac{1}{P(C \geq r|T_1)}\right] Var\left(E[m|T_r]|T_{r-1}\right) |a \leq C \leq b\right],
\end{aligned}$$

which is positive semi-definite by construction. ■

Remark: The results also hold if the moment restrictions in Lemma 7(i) are replaced by:

$$E\left[\frac{I(C=R)}{p_R(T_1)} \frac{p_{[a,b]}(T_1)}{P(a \leq C \leq b)} m(Z;\beta)\right] = 0 \quad \text{and} \quad E\left[\left.\begin{pmatrix} I(C=R) - p_R(T_1) \\ I(a \leq C \leq b) - p_{[a,b]}(T_1) \end{pmatrix}\right| T_1\right] = 0$$

almost surely T_1. This representation is also usable in practice since T_1 is always observed.

B Supplemental Appendix B: Monte Carlo experiment

We will now study the small-sample properties of our proposed estimator EFF and inference based on it for estimands that are similar to those considered in our empirical illustration.

B.1 Simulation design

We will consider a setup reflecting the individual's decision to stay or leave dynamically over periods from programs (e.g., smoking cessation, weight loss), school, job, marriage, experiments, surveys, market, etc. We model this decision to leave after any period as a simple comparison between the individual's expectation of the outcome and their actual outcome after that period. Accordingly, we will consider an R-period program where Y_r is the outcome from staying until the end of the r-th period for $r = 1, \ldots, R$ in the program. We will assume that this outcome is generated as follows. For $t = 1, \ldots, T$, let:

$$Y_t = |Y_{t-1}| + Y_{t-2} + X_t + e_t, \quad \text{where} \quad X_t = X_{t-1} + v_t.$$

e_t and v_t are the model errors.[21] Take X_0, Y_{-1}, Y_0 independently $N(1, 1)$ as the initial state. Our analysis below is not conditional on the initial state, but this could be done. We will take $R = T = 3$, and let X_r be the other observed variables for the r-th period for $r = 1, \ldots, R$.

Let the individual's expectation for the outcome in the r-th period be Y_r^*. Suppose that the individual decides to leave the program at the end of the r-th period, conditional on staying until then, if and only if the actual outcome exceeds the expectation, i.e., $Y_r^* < Y_r$. In other words, let the decision to leave at the end of period r be represented by:

$$I(C = r) \;=\; I(Y_r^* < Y_r) \prod_{j=1}^{r-1} I(Y_j^* \geq Y_j) \quad \text{for } r = 1, \ldots, R-1$$

whereas the decision to never leave be represented by $I(C = R) = 1 - \sum_{r=1}^{R-1} I(C = r)$.

The researcher observes C but not Y_r^*. This means that $Z_1 = (Y_{-1}, Y_0, Y_1, X_{-1}, X_0, X_1)'$,

[21] Estimation of regression coefficients in the case of attrition under some form of MAR in dynamic panel data models with fixed effects has been studied in, e.g., Abrevaya (2019).

$Z_2 = (Y_2, X_2)'$ and $Z_3 = (Y_3, X_3)'$ in our notation. So, the observables are $T_1 = Z_1$, $T_2 = (Z_1', Z_2')'$ and $T_3 = (Z_1', Z_2', Z_3')'$ for those with $C = 1$, $C = 2$ and $C = 3$ respectively.

Our distributional assumptions on the data generating process (DGP) are as follows. e_t and v_t are i.i.d. $N(0,1)$ for all t. $u_r := Y_r^* - Y_r$ is i.i.d. $N(0,7)$ for all r. We stipulate a rather large variance for u_r to abstract away from limited overlap. MAR in (1) is imposed by maintaining that $e_t, v_t, u_r, X_0, Y_{-1}, Y_0$ are mutually independent for all t, r. This results in roughly 62% of the individuals with $C = 1$, 23% with $C = 2$, and 15% with $C = 3$.

There are six different targets $[a,b] = [1,3], [1,1], [2,2], [3,3], [1,2]$ and $[2,3]$ that our theoretical results can accommodate for, and we have simulation results for all of them. For brevity, however, we will focus here on $[a,b] = [1,3], [1,1]$ and $[2,2,]$. ($[3,3]$ is the complete case and is trivial whereas the results for $[1,2]$ and $[2,3]$ are similar to those reported here.)

To define $\beta^0_{[a,b]}$, we take the moment function in (2) as $m(Z;\beta) = Y_3 - \beta$ and consider the three targets $[a,b] = [1,3], [1,1]$ and $[2,2]$ giving three parameters of interest. These target parameters are purposely defined similarly to the estimands in our empirical illustration.

We compute the "true value" of these target parameters numerically by generating data from the above DGP with sample size 10 million, estimating the mean of Y_3 for each subpopulation, and then averaging each mean over 10,000 Monte Carlo trials. Consequently, the three target "true values" are: $\beta^0_{[1,3]} = 9.6162$, $\beta^0_{[1,1]} = 10.5232$ and $\beta^0_{[2,2]} = 8.9914$. As evident from Table 5, the error in this approximation is of a rather small order to seriously affect our subsequent analysis that is based on far smaller (than 10 million) sample size.

Target	Descriptive Statistics					
$[a,b]$ for β	Mean	Std	Median	IQR	Min	Max
$[1,3]$	9.6162	0.0022	9.6162	0.0029	9.6086	9.6249
$[1,1]$	10.5232	0.0027	10.5232	0.0037	10.5111	10.5329
$[2,2]$	8.9914	0.0044	8.9914	0.0060	8.9745	9.0084
$[3,3]$	6.8724	0.0050	6.8724	0.0067	6.8516	6.8924

Table 5: $\beta^0_{[a,b]}$ is approximated (column 2) for different target populations (column 1) based on averaging over 10,000 Monte Carlo trials the target-sample means obtained by using the same DGP and with sample size $n = 10$ million. Columns 3-7 list the standard deviation (Std), interquartile range (IQR), minimum (Min) and maximum (Max) of the estimator.

B.2 Simulation results

We compute our proposed estimator EFF following the description in Section 4. To estimate the nuisance parameters we use as working models the probit models for the conditional hazards and linear models for the conditional expectations. For each working model, we specify the index function as linear in the associated conditioning variables T_1, T_2, etc. and do not include interactions. The true conditional hazards $p^0(.)$'s but not the true conditional expectations $q^0(.)$'s are contained in their respective nuisance working models.

We report in Table 6 the simulation results based on these working models and 10,000 Monte Carlo trials, and for sample size n ranging from quite small to large.[22] We report: (i) Bias, the empirical mean bias; (ii) MC Std, the Monte Carlo standard deviation; (iii) AS Std, the average of the estimated standard error based on the asymptotic variance formula; and (iv) Size, the empirical size of the asymptotic 5% two-sided t test of $H_0 : \beta_{[a,b]} = \beta^0_{[a,b]}$.

Our proposed EFF performs very well in all these aspects (and others) and for all the target $\beta^0_{[a,b]}$ (including those unreported here) even when the sample size n is relatively small.

To put the performance of EFF in context, we also report the same properties of the IPW estimator from (14). IPW performs worse, often much worse, than EFF in every aspect.

First, consider empirical bias. The working parametric models contain the true conditional hazards, i.e., CH holds, and, therefore, IPW and EFF are both asymptotically unbiased. This shows for IPW in the simulation results if we focus on the relatively large samples. On the other hand, the empirical bias of EFF is quite small even in small samples.

Second, consider the variability of the IPW and EFF estimators. MC Std is of course infeasible in practice but is a better measure of the true variability. EFF seems to have much smaller MC Std than IPW. The same observation holds true for AS Std, which is the average of the estimated standard error, a feasible measure, from all the Monte Carlo trials.[23]

[22] $n = 200$ with $P(C = 3) \approx .15$ is small relative to the number of nuisance parameters; $n = 5000$ is not.

[23] We should however note that the observation that MC Std and AS Std are both smaller for EFF than IPW in our simulations is not theoretically promised. This is because: (i) although CH holds, the working models do not contain the true conditional expectations $q^0(.)$'s and hence EFF is not semiparametrically efficient, and (ii) we do not use the Cao et al. (2009)-modification of EFF that, in these cases of scalar

n	Target	Bias		MC Std		AS Std		Size	
	$[a, b]$	EFF	IPW	EFF	IPW	EFF	IPW	EFF	IPW
200	[1,3]	-.043	-.229	.658	1.322	.587	1.006	8.6	15.4
	[1,1]	-.053	-.330	.782	1.580	.728	1.212	7.1	16.9
	[2,2]	-.044	-.132	1.042	1.452	1.003	1.301	6.2	9.0
250	[1,3]	-.030	-.147	.571	1.177	.523	.911	7.4	13.9
	[1,1]	-.042	-.223	.680	1.429	.645	1.105	6.7	15.8
	[2,2]	-.021	-.072	.927	1.313	.897	1.169	6.2	7.8
300	[1,3]	-.024	-.122	.520	1.044	.477	.822	7.8	12.6
	[1,1]	-.032	-.189	.617	1.277	.586	1.004	6.4	14.4
	[2,2]	-.025	-.054	.845	1.172	.816	1.054	6.3	7.4
350	[1,3]	-.021	-.102	.479	.944	.443	.764	7.2	11.7
	[1,1]	-.033	-.160	.566	1.169	.544	.939	6.0	14.2
	[2,2]	-.007	-.040	.782	1.056	.757	.974	6.0	6.7
400	[1,3]	-.014	-.079	.445	.882	.414	.714	6.8	10.8
	[1,1]	-.019	-.125	.530	1.090	.508	.881	5.9	12.6
	[2,2]	-.012	-.033	.730	1.001	.709	.907	5.9	6.5
500	[1,3]	-.020	-.062	.391	.782	.371	.643	6.8	10.1
	[1,1]	-.024	-.093	.472	.988	.454	.800	5.9	11.6
	[2,2]	-.013	-.025	.643	.854	.633	.806	5.3	6.2
750	[1,3]	-.004	-.033	.313	.615	.305	.530	5.4	8.6
	[1,1]	-.008	-.055	.375	.785	.372	.667	4.9	9.9
	[2,2]	-.004	-.014	.522	.681	.518	.650	5.1	5.4
5000	[1,3]	-.002	-.005	.121	.222	.119	.213	5.5	6.3
	[1,1]	-.002	-.008	.145	.290	.145	.276	5.2	6.6
	[2,2]	-.004	-.005	.202	.248	.201	.245	5.1	5.1

Table 6: Results for EFF and IPW are reported based on 10,000 Monte Carlo trials and various sample sizes n. Bias stands for the empirical bias. MC Std and AS Std stands for the standard deviation based on Monte Carlo and the asymptotic variance formula respectively. Size stands for the empirical size of the asymptotic 5% two-sided t-test of $H_0 : \beta_{[a,b]} = \beta^0_{[a,b]}$.

We also note from Table 6 that the feasible measure AS Std resembles very well the infeasible but truer measure MC Std in the case of EFF. Interestingly, on the other hand, AS Std of IPW is much smaller than its MC Std. For practical purpose this means that the user's estimate of the standard error for IPW likely gives a misleadingly higher sense of precision especially in smaller samples. Theoretically, this indicates that the asymptotic approximation better resembles the small sample behavior of EFF than of IPW.

Finally, and extending the discussion of underestimated standard error and quality of asymptotic approximation, we consider Size. Size denotes the empirical size defined as the estimated probability of rejecting the truth by an asymptotic 5% two-sided t test for $H_0 : \beta_{[a,b]} = \beta^0_{[a,b]}$. We observe that Size is much closer to the nominal 5% level for EFF than it is for IPW. (IPW over-rejects the truth much more in small samples.[24]) This is doubly attractive for EFF in these simulations since, as anticipated from our observations on bias and variability, this shows that EFF's gain in precision over IPW comes with another advantage that EFF rejects the truth much less often than IPW, especially in small samples.

Now we move to the case where the nuisance parameters are nonparametrically estimated. The asymptotic variance of IPW estimators should decrease in such cases and, under suitable assumptions, can even reach the efficiency bound; see our Proposition 6(ii) in Section 3.3. Also see, e.g., Hirano et al. (2003), Wooldridge (2007), Chen et al. (2008), Graham (2011), Ackerberg et al. (2014), etc. in similar contexts and Newey (1994), Ackerberg et al. (2012), etc. more generally. We will pursue here this line of argument by obtaining the AS Std of the following three variants of the IPW estimator by enriching the original working model:

- IPW2: based on a working model that augments the original working model (for IPW in Table 6) with the squared terms but no interactions;

parameters of interest, would guarantee that the asymptotic variance of EFF is not larger than that of IPW if CH holds. Nevertheless, it is certainly a welcome observation that EFF delivers estimates that are much more precise than the IPW estimates. We have also noticed this in our other works with more than one level of missingness ($R > 2$). This discussion will need to be modified if the working models "promise" increased flexibility with sample size n; see Ackerberg et al. (2012); and we will do that later with the help of Table 7.

[24] Given Hahn and Liao (2021)'s result of the conservativeness of bootstrap standard error, this observation seems to justify that the anecdotally-common empirical practice of using bootstrap standard errors for IPW.

- IPW2in: based on a working model that augments the original working model (for IPW in Table 6) with the squared terms and all the first order interactions;
- IPW23: based on a working model that augments the original working model (for IPW in Table 6) with the squared and cubic terms but no interactions.

When the progressively richer working models used by these estimators are viewed as a function of sample size n, one would hope that these estimators' asymptotic variances computed as before would eventually converge to the efficiency bound; see, e.g., Newey (1994) and Ackerberg et al. (2012). We report in Table 7 the AS Std and MC Std of IPW2, IPW2in and IPW23 along with IPW and EFF for progressively large sample size. To abstract from: (i) the increased bias (unreported) in smaller samples that is not our focus but nevertheless important and well-studied (see, e.g., Chernozhukov et al. (2018), Rothe and Firpo (2019)) and (ii) more generally from any number of smaller sample issues (see, e.g., Sur and Candes (2019)), we even consider the extremely large sample size of $100,000$.

We also computed another variant IPW23in that is based on a working model that augments the original working model (for IPW in Table 6) with the squared and cubic terms and all first and second order interactions. However, we do not discuss IPW23in except in footnote 26 and omit it from Table 7 because it performs terribly except that when $n = 100,000$, its MC Std is slightly smaller than that of IPW23 (but still bigger, sometimes much bigger, than EFF) that in that instance is the best among the rest of the IPW variants.

We wish to discuss now several observations from Table 7.

First, continuing on the discussion of Table 6, the difference between MC Std and AS Std for each estimator ultimately vanishes with very large sample size ($n = 10,000$ or more).

Second, both MC Std and AS Std of IPW2, IPW2in and IPW23 are smaller than that of IPW for sample size $n = 5000$ and more. This ranking of variability is reassuring since the working models used by IPW2, IPW2in and IPW23 nest the model used by IPW.

Third, although the working models used by both IPW2in and IP23 nest the model used by IPW2, the variability of the former two, as measured by both MC Std and AS Std, seems

n	Target	MC Std					AS Std				
	$[a, b]$	EFF	IPW	IPW2	IPW2in	IPW23	EFF	IPW	IPW2	IPW2in	IPW23
1000	[1,3]	.275	.535	.461	.523	.452	.265	.465	.434	.549	.462
	[1,1]	.329	.687	.579	.633	.549	.322	.589	.539	.667	.569
	[2,2]	.448	.588	.553	.642	.577	.449	.562	.566	.707	.612
5000	[1,3]	.121	.222	.173	.187	.174	.119	.213	.162	.177	.178
	[1,1]	.145	.290	.227	.238	.221	.145	.276	.208	.224	.225
	[2,2]	.202	.248	.220	.235	.225	.201	.245	.220	.238	.235
10000	[1,3]	.085	.157	.113	.116	.114	.084	.152	.110	.114	.118
	[1,1]	.101	.205	.148	.151	.146	.102	.198	.142	.145	.149
	[2,2]	.144	.174	.150	.153	.154	.142	.173	.150	.157	.158
100000	[1,3]	.027	.049	.034	.033	.032	.027	.049	.034	.033	.032
	[1,1]	.033	.064	.044	.042	.040	.032	.064	.044	.042	.041
	[2,2]	.045	.055	.047	.047	.047	.045	.055	.046	.046	.046

Table 7: Standard deviations – MC Std and AS Std – based on Monte Carlo and the asymptotic variance formula respectively of EFF, IPW, IPW2, IPW2in and IPW23 are reported based on 10,000 Monte Carlo trials. The various versions of IPW differ in terms of the specification for the working model for the nuisance parameters – the conditional hazards and expectations, i.e., $p^0(.)$ and $q^0(.)$. In particular, EFF and IPW are those based on the original working model used in Table 6. IPW2 is based on a working model that augments the original working model with the squared terms but no interactions. IPW2in is based on a working model that augments the original working model with the squared terms and all the first order interactions. IPW23 is based on a working model that augments the original working model with the squared and cubic terms but no interactions.

to exceed that of IPW2 even for sample size as large as $n = 10,000$.

The above observations suggest that even in a simple framework such as ours, the sample size of $n = 10,000$ may not be large enough for the intuitions of the large sample theory of IPW to hold convincingly. Other basis functions could lead to a more encouraging picture. Nevertheless, our discussion based on the power series basis is practically relevant since power series resembles the common parametric specification of main variables and interactions used in empirical work and, therefore, it renders the transition from parametric to nonparametric specifications (by adding higher order terms) seamless and empirically palatable.

Fourth, the working models used by IPW2in and IPW23 do not nest each other and hence the ranking of the variability of IPW2in and IPW23 is theoretically unclear. The simulation results lead us to prefer IPW23. For this reason we use this working model in our empirical application (see footnote 14) where the sample size and the dimension of the covariates are comparable to those in the setup here. Some sort of formal regularization or variable selection could be useful, but that is beyond the scope of our current paper.

Finally, we observe from Table 7 that the variabilities, as measured by MC Std and AS Std, of IPW2, IPW2in and IPW23 are still worse, and sometimes much worse, than that of EFF even though EFF is based only on the original working model (as in Table 6).

Let us elaborate on this last observation because this also brings us back to one of our motivations behind extending the MAR analysis to sub-populations with multi-level missingness. To abstract away from any small sample issues that could have worked unfavorably for IPW2, IPW2in and IPW23 because of the large number of nuisance parameters involved in them, let us focus on an extreme case of very large sample size $n = 100,000$.[25]

Now the variabilities of IPW2, IPW2in and IPW23 come quite close to that of EFF for the target $\beta^0_{[2,2]}$. This is a case of only one level of missingness because $R = 3$ while $a = b = 2$; see footnote 9. One level of missingness is what has been considered in the cited references that showed nice properties of IPW based on nonparametric estimation of the conditional

[25] IPW2, IPW2in and IPW23 involve 19, 28 and 55 parameters respectively in their working models for $P(C = 2|C \geq 2, T_2)$ to be estimated based on approximately 38,000 observations ($C \geq 2$) when $n = 100,000$.

hazard (propensity score). Therefore, this closeness of variability and the realization of the promised benefit of nonparametrics is not surprising for the target $\beta^0_{[2,2]}$ when $R = 3$.

However, the variability of IPW2, IPW2in and IPW23 are still substantially larger than that of EFF for the target $\beta^0_{[1,1]}$ that is a case of two levels of missingness since $R = 3$ while $a = b = 1$. We observe the same for $\beta^0_{[1,3]}$ since $\beta^0_{[1,3]}$ is a weighted average involving $\beta^0_{[1,1]}$.[26]

We conclude by restating the three take away points. First, the promises of nonparametrics may not always hold even in very large samples. Second, it is indeed remarkable that the simple EFF estimator fared so well against the other estimators that were based on much richer working models. Third, apart from performing much better than IPW, EFF also performs well in all aspects in absolute terms even in samples of relatively small size.

C Bibliography

Abrevaya, J. (2019). Missing dependent variables in fixed-effects models. *Journal of Econometrics*, 211:151–165.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94:481–498.

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic Efficiency of Semiparametric Two-step GMM. *Review of Economic Studies*, 81: 919–943.

Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96:723–734.

[26] As noted earlier when mentioning IPW23in, enriching the working model further uniformly is not useful here in reducing variability. One could selectively enrich the working model; e.g., enrich it much for $P(C = 1|C \geq 1, T_1)$ since T_1 is observed for all, but keep the model at the level of IPW23 for $P(C = 2|C \geq 2, T_2)$. However, T_1 should be nested in T_2 by the definition of monotonicity. Therefore, selective enrichments that lead to the working model for $P(C = 1|C \geq 1, T_1)$ not being nested in that for $P(C = 2|C \geq 2, T_2)$ are ultimately closer in spirit to imposing parametric restrictions on the MAR assumption. In that case, the target efficiency bounds need to be modified; see, e.g., Hahn (1998) and Chen et al. (2008).

Chaudhuri, S. (2020). On Efficiency Gains from Multiple Incomplete Subsamples. *Econometric Theory*, 36:488–525.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36:808–843.

Chernozhukov, V., Escanciano, J., Ichimura, H., Newey, W., and Robins, J. M. (2018). Locally Robust Semiparametric Estimation. Working Paper.

Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79:437 – 452.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66:315–331.

Hahn, J. and Liao, Z. (2021). Bootstrap Standard Error Estimates and Inference. *Econometrica*, 89: 1963–1977.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores. *Econometrica*, 71:1161–1189.

Newey, W. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62:1349–1382.

Rothe, C. and Firpo, S. (2019). Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. *Econometric Theory*, 35: 1048–1087.

Sur, P. and Candes, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *The Proceedings of the National Academy of Sciences*, 116: 14516–14525.

Wooldridge, J. M. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2):1281–1301.