Instrumental Variable Estimation of Dynamic Treatment Effects on a Duration Outcome

Jad Beyhum, Samuele Centorrino, Jean-Pierre Florens, and Ingrid Van Keilegom

ABSTRACT
This article considers identification and estimation of the causal effect of the time $Z$ until a subject is treated on a duration $T$. The time-to-treatment is not randomly assigned, $T$ is randomly right censored by a random variable $C$, and the time-to-treatment $Z$ is right censored by $T \wedge C$. The endogeneity issue is treated using an instrumental variable explaining $Z$ and independent of the error term of the model. We study identification in a fully nonparametric framework. We show that our specification generates an integral equation, of which the regression function of interest is a solution. We provide identification conditions that rely on this identification equation. We assume that the regression function follows a parametric model for estimation purposes. We propose an estimation procedure and give conditions under which the estimator is asymptotically normal. The estimators exhibit good finite sample properties in simulations. Our methodology could be applied concerns the evaluation of the causal effect of the timing of a reduction in unemployment benefits. The decision to undergo therapy is left to the worker's discretion. The choice of empirical relevance to which our methodology could be applied concerns the evaluation of the causal effect of the timing of a reduction in unemployment benefits on the duration of unemployment of some job seekers.

1. Introduction
In this article, we present a novel econometric methodology for evaluating the causal impact of treatment timing $Z$, on a given duration $T$. Our approach allows the treatment to be administered at any point within the duration spell. To demonstrate the effectiveness of our methodology, we apply it to investigate how the time $Z$ until the start of a therapy for workers on medical leave due to burnout affects the duration $T$ of their leave. Our study uses data from a large Belgian insurance company, which assigns workers to a medical center at the beginning of their leave. The assigned center subsequently evaluates the patients' eligibility for therapy based on the expected medical benefits. The decision to undergo therapy is left to the worker's discretion. Another example of empirical relevance to which our methodology could be applied concerns the evaluation of the causal effect of the timing of a reduction in unemployment benefits on the duration of unemployment of some job seekers.

Within our theoretical framework, we address the issue of endogeneity in the time-to-treatment variable $Z$. In our application, endogeneity may arise because both the medical partner's decision to offer treatment and the worker's decision to accept it are influenced by the patient's unobserved mental health. To tackle this endogeneity problem, we introduce a time-independent instrumental variable, denoted as $W$. Both the time-to-treatment $Z$ and the instrument $W$ are continuous random variables. In our application, the instrument is the proportion of patients assigned to the same medical center who receive treatment within a given year. This serves as a proxy for the availability of the medical center, which is expected to affect the time-to-treatment. The exogeneity of the instrument can be justified by the fact that the treatment center assignment is determined only by geographical factors.

We also allow that the duration of interest $T$ is subject to random right-censoring by a censoring time $C$. This reflects the fact that, in practice, some patients' spells have not yet concluded at the time of data analysis. Moreover, the timing of treatment, $Z$, is itself censored by $T \wedge C$. The censoring of $Z$ by $T$ is endogenous, as these variables are mutually dependent. In our empirical application, this endogenous censoring occurs when workers return to work, rendering them ineligible for treatment.

We establish nonparametric identification of the causal effect of the time-to-treatment on the hazard rate of the outcome duration in our model, relying on two crucial assumptions. The first assumption, known as "no anticipation," implies that subjects' anticipation of treatment timing is unrelated to their actual treatment timing. It allows us to tackle the issue of endogenous censoring of $Z$ by $T$. The second assumption, called "rank invariance," posits that if subject $i$ were to return to work earlier than subject $j$ if both were treated at time $z$, the same ordering would hold if both were treated at a different time, $z' \neq z$. This condition enables us to identify causal effects on the whole population. Leveraging these assumptions, we reframe the duration model as a nonseparable nonparametric instrumental variable (henceforth, NPIV) model, as introduced by Chernozhukov and Hansen (2005). We obtain nonparametric identification of the model by adapting tools from the nonseparable NPIV literature to the context of this article. This
requires us to formalize the relevance of the instrument through completeness conditions.

For estimation purposes, we assume that the regression function follows a parametric functional form, making the model semiparametric. We propose a minimum-distance estimation procedure and provide conditions under which the estimator is asymptotically normal. Additionally, we assess the finite sample properties of the estimator through simulations and illustrate the methodology using the burnout empirical application mentioned earlier.

1.1. Related Literature

This article is first related to a body of work on instrumental variable methods with randomly right-censored duration outcomes, where both the treatment and the instrument are time-independent. We only cite here some references to avoid lengthening the article. Several studies have focused on semiparametric models, such as the additive hazard model Tchetgen et al. (2015), or quantile regression models Chernozhukov, Fernández-Val, and Kowalski (2015). Additionally, other works have provided nonparametric estimation results, including Frandsen (2015), Sant’Anna (2016), Richardson et al. (2017), Blanco et al. (2020), SantAnna (2021), and Beyhum, Flores, and Keilegom (2022). Our article distinguishes itself from this literature because it considers the case where the endogenous variable methods with randomly right-censored durations. However, they do not allow for censoring of the instrument, and therefore the time-to-treatment is either equal to the instrument (that is the predetermined time) or \( \infty \) if the subject chooses not to take the treatment. They identify local average treatment effects on a subset of the population analogous to the compliers in static problems (see Angrist, Imbens, and Rubin 1996). In contrast, the present work does not assume such one-sided noncompliance, and our method allows us to estimate effects over the whole population. Moreover, our empirical application does not fit in their particular one-sided noncompliance framework. Finally, note a recent line of work in biostatistics (Cui et al. 2023; Michael et al. 2023) which studies the case where the instrument varies with time, ruling out the case with a static instrument studied in our article.

1.2. Outline

This article is organized as follows. The model specification is given in Section 2. We study identification in Section 3. Section 4 is devoted to estimation and inference. Section 5 describes our simulations and the empirical application. Concluding remarks are given in Section 6. The proofs of the results and additional numerical experiments are in the online appendix.

2. The Model

2.1. The Duration Model

Let \( T(z) \) be the potential outcome of the duration when the treatment time is set to \( z \in \mathbb{R}_+ = \mathbb{R}_+ \cup \{\infty\} \). When \( z = \infty \), \( T(z) = T(\infty) \) corresponds to the duration that would have been realized if the subject of interest were never treated. The time-to-treatment is a random variable \( Z \), with support \( Z \subseteq \mathbb{R}_+ \). We impose the consistency condition \( T = T(Z) \). Recall that, in our empirical application, the variable \( Z \) is the time until the therapy for burnout starts and \( T \) is the actual duration of medical leave. In this case, \( T(z) \) is the duration of medical leave that would have been realized if the therapy for burnout had been started at time \( z \).

We assume that \( T(z) \) is a continuous random variable and let \( \lambda(z, \cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+ \) be its hazard rate, that is

\[
\lambda(z, t) = \lim_{dt \to 0} \frac{\mathbb{P}(T(z) \in [t, t + dt] | T(z) \geq t)}{dt}.
\]

We call \( \lambda \) the "structural hazard." Remark that \( \lambda(z, \cdot) \) differs from the hazard rate of \( T \) conditional on \( Z = z \) in general (this is the endogeneity issue). The main goal of the article is to identify and estimate this structural hazard.

As mentioned in the introduction, one of the problems considered in this article is that \( Z \) is censored by \( T \). When \( Z > T \), we never know to which potential outcome \( T \) corresponds. To circumvent this issue, we impose the so-called no anticipation assumption (see Abbring and Van den Berg 2003; Van den Berg, Bonev, and Mammen 2020; Van den Berg, Bozio, and Costa Dias 2020), which is standard in the literature on dynamic treatment effects. The no anticipation assumption can be formally stated as follows.

**Assumption 2.1.** For all \( z, z' \in \mathbb{R}_+ \) and \( 0 \leq t < z \land z' \), we have \( \lambda(z, t) = \lambda(z', t) \).
Figure 1. Structural hazard rates under treatment levels $z = 5$ and $z = 7$ obtained by setting $\lambda(5,t) = (5 + 0.2\sqrt{t})(t < 5) + (10 + 0.4\sqrt{t})(5 \leq t)$ and $\lambda(7,t) = (5 + 0.2\sqrt{t})(t < 7) + (6 + 0.6\sqrt{t})(7 \leq t)$.

This assumption means that the hazard rate $\lambda(z,t)$ of $T(z)$ does not depend on $z$ when $z > t$. Under Assumption 2.1, all observations for which $Z > T$ correspond to the same structural hazard (equal to $\lambda(\infty, \cdot)$), which allows us to solve the aforementioned issue that we do not observe $Z$ when $Z > T$. Denote by $\Lambda(z,t) = \int_0^t \lambda(z,s)ds$ the structural cumulative hazard of $T(z)$. By integration and derivation, Assumption 2.1 is also equivalent to $\Lambda(z,t) = \Lambda(z',t)$ when $t < z \wedge z'$, which is the way the no anticipation assumption is stated in Abbring and Van den Berg (2003) (except that the strict inequality $t < z \wedge z'$ is replaced by the weak inequality $t \leq z \wedge z'$ in Abbring and Van den Berg 2003). We illustrate the no anticipation assumption in Figure 1, where we draw the structural hazard rates under time-to-treatment equal to $z = 5$ and $z = 7$ (in the caption, we use $I(\cdot)$ to denote the indicator function). In this illustration, the treatment increases the hazard rate, but our model also allows the treatment to reduce the hazard rate. Before the treatment, the hazard rate does not depend on the treatment time, while after the treatment it does.

The no anticipation assumption does not rule out that subjects actually anticipate receiving future treatment. It rather means that the way agents anticipate future treatment cannot be related to their actual time-to-treatment (see the discussion in Abbring and Van den Berg 2003 for more details). In the application it can be justified by the fact that workers are not given any information on when they will be evaluated by the medical partner.

To be able to identify treatment effects over the full population under endogeneity, we impose constraints on the unobserved heterogeneity of the model. For $z \in \mathbb{R}_+$ and $t \in \mathbb{R}_+$, recall that $\Lambda(z,t) = \int_0^t \lambda(z,s)ds$ is the structural cumulative hazard under treatment $z$ and let $U(z) = \Lambda(z,T(z))$ be the hazard of $T(z)$ evaluated at $T(z)$. The $\{U(z)\}_{z \in \mathbb{R}_+}$ can be thought of as the unobserved heterogeneity of the model. In our empirical application, it could represent the underlying mental health of the subject. We impose the following conditions on them.

Assumption 2.2. The following holds:

(i) There exists a random variable $U$ such that $U(z) = U$, for all $z \in \mathbb{R}_+$;

(ii) For all $z \in \mathbb{R}_+$, $\lambda(z, \cdot)$ is continuous on $[0,z)$ and $[z, \infty)$.

Condition (ii) allows the hazard rate $\lambda(z, \cdot)$ to be discontinuous at the time of treatment $z$. This permits behaviors similar to that of Figure 1, where treatment makes the hazard rate “jump.” We show in the online appendix (Lemma S.1.1) that Assumption 2.2 (ii) implies that the cumulative hazard $\Lambda(z, \cdot)$ maps the support of $T(z)$ to $\mathbb{R}_+$ and is strictly increasing on the support of $T(z)$. Therefore, we can define $\Lambda(z, \cdot)^{-1}$, the inverse of the mapping $\Lambda(z, \cdot)$ restricted to the support of $T(z)$ (that is $\Lambda(z, \cdot)^{-1}(u)$ is the unique element $t$ of the support of $T(z)$ such that $\Lambda(z,t) = u$). This implies that $T(z) = \Lambda(z, \cdot)^{-1}(U(z))$ is strictly increasing in $U(z)$. Then, Assumption 2.2 (ii) yields that, for two subjects $i$ and $j$, $T_i(z) > T_j(z)$ implies $U_i(z) = U_i > U_j = U_j(z)$, which leads to $T_i(z') > T_j(z')$, for all $z,z' \in \mathbb{R}_+$. In other words, Assumption 2.2 (i) implies that the rank in the outcome of any two subjects is the same across all potential outcomes. Assumption 2.2 (i) is therefore a rank invariance assumption as in Chernozhukov and Hansen (2005). This assumption restricts the heterogeneity of the treatment effects on the duration: the treatment can change the quantiles of the distribution of the potential outcomes but it cannot change the rank that a subject has in this distribution. Moreover, remark that the rank invariance assumption does not restrict the possible values of the structural hazard $\lambda(z,t)$ beyond the continuity condition in Assumption 2.2 (ii) and therefore, in this sense, rank invariance is not a constraint on the marginal distribution of $T(z)$. It only imposes limits on the joint distribution of potential outcomes, that is the distribution of $(T(z))_{z \in \mathbb{Z}}$. Note also that we could relax the rank invariance assumption into a rank similarity assumption as in Chernozhukov and Hansen (2005) while keeping all results valid.

Finally, in the context of the empirical application, this rank invariance assumption means that if subject $i$ were to return to work earlier than subject $j$ if both were treated at time $z$, then $i$ would also start working again before $j$ if both were treated at $z' \neq z$.

Let us now discuss useful consequences of Assumption 2.2.

A first consequence is that $\Lambda(Z,T) = U$. This relation will be helpful to reformulate the model as a nonseparable NPIV model later in the current section. Moreover, the cumulative hazard of a duration applied to this duration follows a unit exponential distribution (denoted $\text{Exp}(1)$), so that for all $z \in \mathbb{R}_+$, $\Lambda(z,T(z)) = U(z) \sim \text{Exp}(1)$ (see Lemma S.1.1 in the online appendix for a formal proof). Hence, it also follows from Assumption 2.2 (i) that $U \sim \text{Exp}(1)$. This fact will be useful to derive identification results. It is important to note that $U \sim \text{Exp}(1)$ is not an assumption in itself, but rather a consequence of the definition of $U(z)$ and rank invariance. Note that in the nonseparable NPIV model of Chernozhukov and Hansen (2005), the error term follows a standard uniform distribution. In the present article, $U$ instead follows a standard exponential distribution which is more natural in duration models.

When $U$ and $Z$ are independent, $\lambda(z, \cdot)$ is the hazard rate of $T$ given $Z = z$. However, in practice, the variables $U$ and $Z$ may be dependent, which creates an endogeneity issue, that is $\lambda(z, \cdot)$ may differ from the hazard rate of $T$ given $Z = z$. In the context of the empirical application, the endogeneity may, for instance, be due to the fact that subjects with worse burnout are
more likely to be treated early. There exists an instrument \( W \) allowing to solve this issue. The support of \( W \) is denoted \( \mathcal{W} \). For simplicity, we limit ourselves to the case where \( W \) is scalar, that is \( \mathcal{W} \subset \mathbb{R} \). We impose the following assumption:

**Assumption 2.3.** \( W \) is independent of \( U \).

Recall that, in the empirical application, the instrument corresponds to the proportion of patients assigned to the same medical center who receive treatment within a given year. The workers are assigned to the treatment center the closest to their workplace. Hence, Assumption 2.3 can be interpreted as the fact that the location of the workplace is independent of the underlying mental health of the patient.

The duration \( T \) is randomly right censored by a random variable \( C \) with support in \( \mathbb{R}_+ \), so that we do not observe \( T \) but \( Y = T \wedge C \). The observables are \((Y, \delta, \tilde{Z}, \tilde{D}, W)\), where \( \delta = I(T \leq C), \tilde{Z} = Z \wedge Y \), is a censored version of \( Z \) and \( \tilde{D} = I(Z \leq Y) \) is a censored version of \( D = I(Z \leq T) \), the treatment indicator. Note that \( D = 1 \) for treated observations only (in the sense that they receive treatment before the end of their spell). In the burnout data, \( C \) corresponds to the duration during which subjects on medical leave are followed in the data (the follow-up stops after two years of medical leave or at the end of 2020), we have \( \delta = 1 \) if the duration of medical leave is observed and 0 otherwise, \( \tilde{Z} \) is the minimum between the treatment time, the censoring time and the duration of medical leave, \( D = 1 \) if the subject is treated before the end of its medical leave and 0 otherwise and \( \tilde{D} = 1 \) if we observe the treatment time in the dataset and 0 otherwise.

### 2.2. Reformulation as a Nonseparable NPIV Model

As mentioned in the introduction, this article makes use of tools from the NPIV model to solve the dynamic problem. Recall that, for all \( z \in \mathbb{R}_+ \), \( \Lambda(z, \cdot)^{-1}(u) \) is the unique element \( t \) of the support of \( T(z) \) such that \( \Lambda(z, t) = u \) and \( I(\cdot) \) denote the indicator function. We can reformulate the model as a nonseparable NPIV model by inverting the relation \( \Lambda(Z, T) = U \). This is formalized in the following lemma.

**Lemma 2.1.** Under Assumptions 2.1 and 2.2, we can write

\[
T = \varphi(Z, U) = I(Z > \varphi_0(U))\varphi_0(U) + I(Z \leq \varphi_0(U))\varphi_1(Z, U) \quad (1)
\]

where \( \varphi_0 : \mathbb{R}_+ \mapsto \mathbb{R}_+ \) is equal to \( \Lambda(\infty, \cdot)^{-1} \), \( \varphi_1(z, \cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+ \) is equal to \( \Lambda(z, \cdot)^{-1} \). Moreover, for all \( z \in \mathbb{R}_+ \), we have \( \varphi_1(z, \varphi_0^{-1}(z)) = z \).

Equation (1) means that there exists mappings \( \varphi_0 \) and \( \varphi_1 \) such that \( T \) is equal to \( \varphi_0(U) \) when \( Z > \varphi_0(U) \) and equal to \( \varphi_1(Z, U) \) otherwise. Intuitively, this signifies that there are two regressions functions: one of the “not yet treated” corresponding to \( \varphi_0 \), and one for the “already treated” which is \( \varphi_1 \). Writing the NPIV model as functions of these \( \varphi_0 \) and \( \varphi_1 \) allows us to define a parameter space which is a vector space (see Section 3.2).

Note that, by definition, \( \varphi_0(\cdot) \) and \( \varphi_1(z, \cdot) \) are strictly increasing. Moreover, the fact that \( \varphi_1(z, \varphi_0^{-1}(z)) = z \) for all \( z \in \mathbb{R}_+ \) (see the end of Lemma 2.1) implies that \( \psi \) is strictly increasing too.

Equation (1) defines a nonseparable NPIV model similar to that of Chernozhukov and Hansen (2005). Hence, we can use tools from the literature on nonseparable NPIV models to obtain identification results on \( \varphi \). Throughout the article, we will therefore identify and estimate \( \varphi \) rather than \( \lambda \) because this approach simplifies the mathematical analysis. It follows from Assumption 2.1 that \( \varphi(z, \cdot) = \Lambda(z, \cdot)^{-1} \). Hence, \( \lambda(z, t) \) is the derivative of the inverse of \( \varphi(z, \cdot) \) at \( t \). As a result, \( \psi \) is a one-to-one transformation of \( \lambda \) and identification of \( \psi \) implies identification of \( \lambda \) and therefore of many quantities of interest in duration models (such as the survival function or the cumulative hazard). Note that, the present model involves two additional complexities with respect to the original model of Chernozhukov and Hansen (2005).

As we will see later, this implies that we have to adapt results of the nonseparable NPIV model to our nonstandard model.

### 3. Identification

We study identification in a fully nonparametric setting. First, we show that our model generates a system of integral equations. Then we derive identification results based on this system of equations.

#### 3.1. Identification Equation

In order to formulate the identification equation, we introduce the following reduced form quantities. For \( t \in \mathbb{R}_+ \), \( w \in \mathcal{W} \), let

\[
F_0(t, w) = \mathbb{P}(T \leq t, D = 0, W \leq w); F_W(w) = \mathbb{P}(W \leq w).
\]

Moreover, for a mapping \( \phi : \mathcal{Z} \mapsto \mathbb{R}_+ \) and \( w \in \mathcal{W} \), we define

\[
F_1(\phi, w) = \mathbb{P}(T \leq \phi(Z), D = 1, W \leq w).
\]

The following theorem states the system of equations that we use to obtain identification results.

**Theorem 3.1.** Let Assumptions 2.1, 2.2, and 2.3 hold. For the true \( \varphi_0 \) and \( \varphi_1 \) (defined below (1)) and all \( w \in \mathcal{W} \), we have

\[
F_0(\varphi_0(u), w) + F_1(\varphi_1(\cdot, u), w) = (1 - e^{-k})F_W(w). \quad (2)
\]

Equation (2) essentially follows from the independence of \( U \) and \( W \) and the fact that \( U \) follows a unit exponential distribution. In Section S.1 of the online Appendix, we provide an alternative characterization of the model in terms of reduced-form (conditional) hazard rates and survival functions of the observed duration \( T \), which are natural quantities in the context of duration models.

#### 3.2. Identification Without Censoring

In this section, we discuss the simple case where there is no censoring, that is \( C = \infty \) a.s. Then, \( Y = T \) and \( \tilde{D} = D \), which implies that \( F_0(t, w) \) is identified for all \( t \in \mathbb{R}_+ \) and \( w \in \mathcal{W} \).
Moreover, when \(D = 1\), we have \(\tilde{Z} = Z\) and, hence, \(F_1(\phi, w)\) is identified for all \(\phi : Z \mapsto \mathbb{R}_+\) and \(w \in W\). Therefore, \(F_0, F_1, F_W\) in (2) are all identified in the absence of censoring. Hence, uniqueness of the solutions to (2) implies identification. This allows us to derive identification results in the next two sections.

We focus on identification of \(\varphi(\cdot, u)\) for a given \(u \in \mathbb{R}_+\). At this point, it is useful to define the set to which \((\varphi_0(u), \varphi_1(\cdot, u))\) belongs. Let the parameter space \(\mathcal{P}\) be the set of \((\psi_0, \psi_1)\) such that \(\psi_0 \in \mathbb{R}\), and \(\psi_1\) is a bounded mapping from \(Z\) to \(\mathbb{R}\). This set \(\mathcal{P}\) is a vector space and we endow it with the norm \(\|\cdot\|_P\), where \(\|\psi_0, \psi_1\|_P = E[\psi(Z)^2]U = u\), with \(\psi(z) = \psi_0(z > \psi_0) + \psi_1(z)(z \leq \psi_0)\). The function \(\psi\) is the mapping “induced” by \((\psi_0, \psi_1)\). It is similar to the object that we want to identify \((\varphi(\cdot, u))\). The norm \(\|\psi_0, \psi_1\|_P\) is finite because \(\psi_1\) is bounded. We assume that \(\psi_1\) is bounded, because, in practice, to identify \(\varphi_1(\cdot, u)\) in the presence of censoring with finite support, we will have to assume that \(\varphi_1(\cdot, u)\) is bounded by the upper bound of the support of the censoring variable (see the discussion in Section 3.3 and Assumption 4.1(ii) later in the article). The fact that we restrict the parameter space to bounded functions \(\psi_1\) also allows us to weaken some of the conditions for identification.

Remark also that since (i) \(\mathcal{P}\) is not a standard \(L^2\)-space with respect to a continuous distribution and (ii) what we wish to identify is not \((\varphi_0(u), \varphi_1(\cdot, u))\) but \(\varphi(\cdot, u)\), we cannot rely on the high-level identification theory for nonseparable NPIV models as in Chernozhukov and Hansen (2005) and Chen et al. (2014). Therefore, to obtain our identification results, we adapt the proofs of these papers to our specific model.

### 3.2.1. Local Identification

We start by local identification.

**Definition 3.1.** The regression function \(\varphi(\cdot, u)\) is locally identified in a set \(\mathcal{N} \subset \mathcal{P}\) for all \((\psi_0, \psi_1) \in \mathcal{N}\),

\[
F_0(\psi_0, w) + F_1(\psi_1, w) = (1 - e^{-\varepsilon u})F_W(w),
\]

implies that \(\psi(Z) = \varphi(Z, u)\) almost surely.

Notice that we only seek to identify \(\varphi(z, u)\) for all \(z \in Z\) and not \((\varphi_0(u), \varphi_1(z, u))\). This is because it is not possible (and not interesting) to identify \(\varphi_1(z, u)\) because \(\varphi_1(z, u)\) never generates the data for such \(z\). The main assumption for local identification is the following bounded completeness condition.

**Assumption 3.1.** For all \((\psi_0, \psi_1) \in \mathcal{P}\),

\[
E[\psi_0I(Z > \psi_0)] + I(Z \leq \psi_0)\psi_1(Z)|U = u, W = 0 \text{ a.s.} \Rightarrow P(\psi_0I(Z > \psi_0) + I(Z \leq \psi_0)\psi_1(Z) = 0|U = u) = 1.
\]

Intuitively, this condition means that \(Z\) and \(W\) are sufficiently dependent given \(U = u\). **Assumption 3.1** is implied by the bounded completeness of \(Z\) given \(U = u\), that is for all bounded functions

\[
m : Z \mapsto \mathbb{R}, E[m(Z)|U = u, W] = 0 \text{ a.s.} \Rightarrow P(m(Z) = 0|U = u) = 1.
\]

Such a condition is imposed in Cazals et al. (2016). As noted by a reviewer, (4) means that there exists no nonzero function of \(Z\) which has mean zero given \(W, U = u\). Moreover, note that condition (4) is unrelated to \(\lambda\) and, therefore, does not depend on assumptions made on it but solely on the joint distribution of \(Z, W\) given \(U = u\). In separable NPIV models, the related bounded completeness condition

\[
m : Z \times \mathbb{R}_+ \mapsto \mathbb{R}, E[m(Z)|U = u, W] = 0 \text{ a.s.} \Rightarrow m(Z) = 0 \text{ a.s.} \quad (5)
\]

for \(m\) belonging to some class of functions, is often imposed (see Newey and Powell 2003, Darolles et al. 2011, among others). Condition (4) is a “conditional on \(U = u\)” version of condition (5). Several authors have provided sufficient conditions for bounded completeness as stated in (5) (see Newey and Powell 2003, D’Haultfoeuille 2011, Hu and Shiu 2018, Andrews 2017, among others). These sufficient conditions are restrictions on the family of distributions \([Z|W = w]\), where \([Z|W = w]\) stands for the distribution of \(Z\) given \(W = w\). To obtain sufficient conditions for (4), it therefore suffices to impose the sufficient conditions from Newey and Powell (2003), D’Haultfoeuille (2011), Hu and Shiu (2018), Andrews (2017) on the family of distributions \([Z|W = w, U = u]_w\) rather than \([Z|W = w]_w\). As an illustration, we adapt the conditions of Newey and Powell (2003) to show condition (4) in Section S.1.5.4 of the online appendix. Notice also that restricting conditions (4) to bounded functions makes it more likely to hold (see the aforementioned papers for further details).

In the context of the empirical application, **Assumption 3.1** means that \(W\), the proportion of workers in the same medical center who are offered the treatment, and the time-to-treatment \(Z\) are sufficiently dependent. The variables \(Z\) and \(W\) are likely to be dependent because centers with more availability are both likely to treat a higher proportion of patients (large value of \(W\)) and offer treatment earlier (small value of \(Z\)).

In addition to **Assumption 3.1**, we also impose some regularity conditions. Since these conditions are technically involved, they are stated in the online appendix (see Assumption S.1 in Section S.1.5). Remark that these regularity conditions require Gâteaux differentiability of some operator but no Fréchet differentiability is needed. We have the following local identification result.

**Theorem 3.2.** Let Assumptions 2.1, 2.2, 2.3, 3.1, S.1 hold and assume that \((\varphi_0(u), \varphi_1(\cdot, u)) \in \mathcal{P}\). Then, for all \((\psi_0, \psi_1) \in \mathcal{P}\), there exists \(\varepsilon > 0\) such that \(\varphi(\cdot, u)\) is locally identified on

\[
\mathcal{N}_\varepsilon = \{(\varphi_0(u) + \varepsilon \psi_0, \varphi_1(\cdot, u) + \varepsilon \psi_1) : \varepsilon \in [-\varepsilon, \varepsilon]\}
\]

We have shown here local identification on a segment \(\mathcal{N}_\varepsilon\). As noted in Chen et al. (2014), in nonparametric nonlinear structural models (as the one of the present article), it is often not possible to derive local identification results when \(\mathcal{N}\) is an open ball (in the topology defined by \(\|\cdot\|_P\)). We therefore focused on a smaller set, which is not an open ball.

### 3.2.2. Global Identification

Next, we discuss global identification, which we define as follows.
**Definition 3.2.** The function \( \psi(\cdot, u) \) is globally identified if for all \((\psi_0, \psi_1) \in \mathcal{P} \), the fact that (3) holds implies that \( \psi(Z) = \varphi(Z, u) \) almost surely.

We adapt the theory of Chernozhukov and Hansen (2005) to the case of the present model. We introduce \( \varepsilon = T - \varphi_0(u)(1 - D) - \varphi_1(Z, u)D \). Let \( f_{E[D,W]}(-|0, w) \) be the density of \( \varepsilon \) given \( D = 0 \) and \( W = w \), \( f_{E[D,Z,W]}(1, z, w) \) be the density of \( \varepsilon \) given \( D = 1, Z = z \) and \( W = w \) and \( f_{E[D]}(d) \) be the density of \( Z \) given \( D = d \). (Their existence is guaranteed under our assumptions). Let us make the following Assumption:

**Assumption 3.2.** The distribution of \((U, Z, W)\) is absolutely continuous with continuous density, \(0 < P(D = 1) < 1\), and there exists a constant \( K > 0 \) such that \( f_{E[D]}(d) = 0 \) for all \( d \in \mathbb{R}_+ \), \( f_{E[D,Z,W]}(1, z, w) = 0 \) for all \( z \in \mathbb{R}_+ \) and \( w \in \mathbb{R}_+ \) and \( f_{E[D]}(d) = 0 \) for all \( d \in \mathbb{R}_+ \) (their existence is guaranteed under our assumptions).

For \((\Delta_0, \Delta_1) \in \mathcal{P}\), we define

\[
\omega_\Delta(Z, D, W) = \left[ \int_0^1 f_{E[D,W]}(\frac{\varepsilon}{\Delta_0})|0, W) d\varepsilon \right] (1 - D) + \left[ \int_0^1 f_{E[D,Z,W]}(\frac{\varepsilon}{\Delta_1})|1, Z, W) d\varepsilon \right] D.
\]

We make the following hypothesis:

**Assumption 3.3.** For all \((\Delta_0, \Delta_1) \in \mathcal{P}\), we have

\[E[(\Delta_0(1 - D) + \Delta_1(1 - D))\omega_\Delta(Z, D, W)|W] = 0 \text{ a.s.}\]

\[\Rightarrow \Delta_0(1 - D) + \Delta_1(1 - D) = 0 \text{ a.s.}\]

This is a type of bounded strong completeness condition. It is the counterpart of Assumption L1* of Chernozhukov and Hansen (2005) in our model. This assumption holds when \( Z \) and \( W \) are sufficiently dependent given \( U \). To the best of our knowledge, the only sufficient conditions known for this type of assumption correspond to condition L2* in Chernozhukov and Hansen (2005) (see Canay, Santos, and Shaikh (2013) for a discussion regarding the lack of simple identification conditions in nonlinear models with endogeneity such as ours). In Section S.1.6.3 of the online appendix, we give sufficient conditions for Assumption 3.3 which are in the spirit of condition L2* in Chernozhukov and Hansen (2005). These conditions include assuming that a family (in \( w \)) of distributions is bounded complete and therefore relate Assumptions 3.3 to standard bounded completeness conditions (as in (5)). We have the following global identification result.

**Theorem 3.3.** Let Assumptions 2.1, 2.2, 2.3, 3.2, 3.3 hold and assume that \((\psi_0(u), \psi_1(\cdot, u)) \in \mathcal{P}\), then \( \psi(\cdot, u) \) is globally identified.

### 3.3. Identification with Censoring

Let us now consider the case where \( T \) is right censored. In this case, \( F_0 \) and \( F_1 \) may not be identified everywhere because of censoring. We make the following assumption on the censoring:

**Assumption 3.4.** The censoring time \( C \) is independent of \((U, Z, W)\). This assumption could be relaxed. For instance, the fact that \( C \) and \( U \) are independent given \( Z, W \) would suffice for identification. However, we impose the stronger Assumption 3.4 to simplify the exposition and the estimation.

Let \( c_0 \) be the upper bound of the support of \( C \). Identification of \( F_0(t, w) \) and \( F_1(\psi, w) \) for \( \psi : Z \mapsto [0, t] \) is only possible for \( t \in [0, c_0] \). Hence, we can only check that \( \psi(\cdot, u) \) is the solution of the identification equation for \( u \in [0, u_0] \), where \( u_0 = \sup u \in [0, c_0] \), the survival function of \( C \). Since \( F_0(t, w) = E[I(T \leq t, D = 0, W \leq w)] \), using the law of iterated expectations and Assumption 3.4, we can show that

\[
F_0(t, w) = E\left[ \delta_{G(Y)}I(Y \leq t, D = 0, W \leq w) \right],
\]

for all \( t \in [0, c_0] \). Similarly, for all \( \phi : Z \mapsto [0, c_0] \), it holds that

\[
F_1(\phi, w) = E\left[ \delta_{G(Y)}I(Y \leq \phi(\tilde{Z}), D = 1, W \leq w) \right].
\]

The proof of (6) and (7) is given in Section S.1.1 of the online appendix. By standard arguments from the survival analysis literature, \( G(t) \) is identified for all \( t \in [0, \sup u \in \mathbb{R}_+ : P(T \geq t) > 0] \) (on this interval \( G(t) \) is equal to the population analog of the Kaplan-Meier estimator of the survival function of \( C \) which identifies it). Hence, the quantities on the right-hand side of (6) and (7) are identified and so are \( F_0(t, w) \) for all \( t \in [0, c_0] \) and \( F_1(\phi, w) \) for all \( \phi : Z \mapsto [0, c_0] \). Therefore, identification results on \( \psi(\cdot, u) \) for \( u \in [0, u_0] \) can be obtained as in the case without censoring. Note however that the fact that we can identify \( \psi \) only up to \( u_0 \) has several implications.

First, it means that average treatment effects \( E[I(T \leq \phi(\tilde{Z}) - T(\phi))] \) are not identified. Second, only some quantile treatment effects \( \psi(z, \cdot) - \psi(z', u) \) are identified. Third, the structural hazard \( \lambda(z, t) = (\varphi(\tilde{Z}) - 1)'(t) \) is identified only for \( t \leq \psi(z, u_0) \).

### 4. Estimation

#### 4.1. Parametric Regression Function

Our strategy is to first estimate \( F_0, F_1, \) and \( F_W \), and then to solve an estimate of (2) for \( \psi_0 \) and \( \psi_1 \), which is obtained by plug-in of the estimates of \( F_0, F_1, \) and \( F_W \). Since (2) is a complicated integral equation, this method is unlikely to deliver precise nonparametric estimates of \( \psi \) on datasets of reasonable size. As a result, we decide to assume that \((\psi_0, \psi_1)\) follows a parametric model \( \psi_{\theta_0, \theta_1} \) for some \( \theta \in \Theta \), where \( \Theta \subset \mathbb{R}^K \) is the parameter set. Here, for all \( \theta \in \Theta \), \( \psi_{\theta_0, \theta_1} \) is a mapping from \( \mathbb{R}_+ \) to \( \mathbb{R}_+ \) and \( \psi_{\theta_1} \) is a mapping from \( Z \times \mathbb{R}_+ \) to \( \mathbb{R}_+ \). For all \( \theta \in \Theta \), we can also define \( \psi_{\theta} : Z \times \mathbb{R}_+ \mapsto \mathbb{R}_+ \) such that \( \psi_{\theta}(z, u) = \psi_{\theta_0, \theta_1}(u)(z > \psi_{\theta_0}(u)) + \psi_{\theta_1}(z, u)(z \leq \psi_{\theta_0}(u)) \) for all \( z \in Z, u \in \mathbb{R}_+ \). This approach has three additional advantages. First, it avoids the need for regularization, since the parameter set has finite dimension. Second, parametric shapes enable to summarize simply the properties of \( \psi \). Finally, they allow to estimate \( \psi(\cdot, u) \) even for \( u > u_0 \), which is useful when the study has insufficient follow-up. Note that, the model remains semiparametric since \( F_0 \) and \( F_1 \) are not parametrically constrained. Let us give some examples of parametric models for \( \psi \).
Example 1 (Weibull model). A first example comes from the Weibull distribution. Recall that \( T(z) \) is the potential outcome of \( T \) when the treatment time is set to \( z \). We assume that before \( z \), the hazard rate of \( T(z) \) corresponds to that of a Weibull distribution with parameters \( \theta_0, \theta_1 \). After \( z \), the hazard rate \( T(z) \) is that of a Weibull distribution with parameters \( \theta_0, \theta_1 \). The structural hazard of \( T(z) \) at time \( t \) is therefore given by

\[
\lambda(z, t) = \theta_0 \theta_1 t^{\theta_1 - 1} I(t < z) + \theta_0 \theta_1 t^{\theta_1 - 1} I(t \geq z).
\]

By inverting the cumulative hazard, it can be shown that

\[
\psi_0(u) = \left( \frac{u}{\theta_0} \right)^{1/\theta_1}; \quad \psi_1(z, u) = \left( \frac{u - \theta_0 z^{\theta_0}}{\theta_1} + z^{\theta_1} \right)^{1/\theta_1}.
\]

Example 2 (Log-normal model). The second example comes from the log-normal distribution. Before \( z \), the hazard rate of \( T(z) \) is assumed to be equal to that of a log-normal distribution with mean \( \theta_0 \) and variance \( \theta_1^2 \). After \( z \), the hazard rate \( T(z) \) corresponds to that of a log-normal distribution with mean \( \theta_1 \) and variance \( \theta_1^2 \). The structural hazard of \( T(z) \) at time \( t \) is then given by

\[
\lambda(z, t) = \theta_0 \theta_1 t^{\theta_1 - 1} I(t < z) + \theta_1 t^{\theta_1 - 1} I(t \geq z),
\]

where \( \phi \) and \( \Phi \) are respectively the density and the cumulative distribution function of a standard normal distribution. Inverting the cumulative hazard, we obtain

\[
\psi_0(u) = \exp \left[ \theta_0 + \theta_1 \Phi^{-1} \left( 1 - \exp(-u) \right) \right],
\]

\[
\psi_1(z, u) = \exp \left[ \theta_1 + \theta_1 \Phi^{-1} \left( 1 - \exp\left(-u + \log(R_z(\theta))\right) \right) \right],
\]

where \( R_z(\theta) = \left[ 1 - \Phi \left( \frac{\log(z) - \theta_0}{\theta_1} \right) \right] \left[ 1 - \Phi \left( \frac{\log(z) - \theta_0}{\theta_1} \right) \right]^{-1} \).

4.2. Estimation of the Integral Equation

For \( \theta \in \Theta \), let

\[
M_0(u, w) = F_0(\psi_0(u), w) + F_1(\psi_1(\cdot, u), w) - (1 - e^{-w})F_W(w)
\]

be the value of the identifying \( (2) \) in \( (\psi_0, \psi_1) \). In order to estimate \( \hat{\psi}_u \) using \( (2) \), it is necessary to estimate the unknown operator \( M \). Assume that we possess an iid sample \( \{Y_i, \delta_i, Z_i, D_i, W_i\}_{i=1}^n \). We estimate \( F_0 \) and \( F_1 \) using \( (6) \) and \( (7) \). Let

\[
N(t) = \sum_{i=1}^n I(Y_i \leq t, \delta_i = 0); \quad Y(t) = \sum_{i=1}^n I(Y_i \geq t).
\]

The Kaplan-Meier estimator of \( G(t) \) is given by \( \hat{G}(t) = \prod_{s \leq t} \left( 1 - \frac{dN(s)}{Y(s)} \right) \), where \( dN(s) = N(s) - \lim_{s' \to s^+} N(s') \). In turn, \( F_0 \) and \( F_1 \) are estimated by

\[
\hat{F}_0(t, w) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{G(Y_i)} I(Y_i \leq t, \delta_i = 0, W_i \leq w);
\]

\[
\hat{F}_1(\psi, w) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{G(Y_i)} I(Y_i \leq \psi(\hat{Z}_i), \hat{D}_i = 1, W_i \leq w).
\]

Then, \( F_W \) is estimated by \( \hat{F}_W(w) = n^{-1} \sum_{i=1}^n I(W_i \leq w) \). Finally, the estimator of \( M \) is

\[
\hat{M}_0(u, w) = \hat{F}_0(\psi_0(u), w) + \hat{F}_1(\psi_1(\cdot, u), w) - (1 - e^{-w})\hat{F}_W(w).
\]

Remark that, although \( Z \) and \( W \) are continuous random variables, we avoid smoothing because we use an unconditional identification equation.

4.3. Estimator of \( \theta_u \)

It is computationally impossible to solve the estimated identifying equation for every \( u \) and \( w \). Hence, we solve it on a grid \( 0 \leq u_1 < \cdots < u_m < u_0 \) of values of \( u \), where \( m \in \mathbb{N} \) is fixed, and at \( w = W_1, \ldots, W_n \). The estimator of \( \theta_u \) is

\[
\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{L}(\theta),
\]

where

\[
\hat{L}(\theta) = (nm)^{-1} \sum_{i=1}^n \sum_{j=1}^m p(u_j) \hat{M}_0(u_j, W_i^2)
\]

is some weighting function, typically \( p(u) = e^{-u} \). This type of estimator is akin to minimum distance from independence estimators as in Brown and Wegkamp (2002). We can not rely directly on the theory of Brown and Wegkamp (2002). Indeed, in our case, because of censoring, \( G \) has to be estimated in a first step. Moreover, weights on \( W \) are chosen according to the empirical measure, while they are set according to a fixed measure selected by the researcher in Brown and Wegkamp (2002). This avoids the need to choose weights on \( W \) and might lead to greater efficiency.

4.4. Asymptotic Normality

Now, we state the conditions that we impose to show asymptotic normality. The first assumption ensures identification, that is

\[
\theta_u = \arg \min_{\theta \in \Theta} L(\theta),
\]

where

\[
L(\theta) = m^{-1} \sum_{j=1}^m p(u_j) E[M_0(u_j, W_j^2)].
\]

Assumption 4.1. The following holds

(i) The regression function \( \psi(\cdot, u_j) \) is globally identified for all \( j = 1, \ldots, m \).
(ii) We have \( \sup_{\theta \in \Theta, z \in Z} \psi_0(u_m) \cap \psi_1(z, u_m) < c_0 \).
(iii) For all \( \theta, \hat{\theta} \in \Theta \) such that \( \psi_0(u_j, z) = \psi_\theta(z, u_j) \), for all \( j \in \{1, \ldots, m\} \), \( z \in Z \), we have \( \theta = \hat{\theta} \).

Condition (i) was studied in the previous section, whereas condition (ii) restricts the choices of \( \Theta \) and \( u_m \), and (iii) is a constraint on the parametric family that guarantees that \( \theta_u \) is identified when \( \psi(\cdot, u) \) is known for all \( j = 1, \ldots, m \). Together (i) and (iii) ensure that \( \theta \) is identified from \( (2) \), which yields \( (11) \).

Condition (ii) implies that \( \psi_{0,1}(\cdot, u) = \psi_{\theta}(\cdot, u) \) is bounded when \( c_0 < \infty \) (that is censoring has finite support). Let \( \| \cdot \| \) denote the Euclidean norm in \( \mathbb{R}^K \). We also impose some regularity conditions:
Theorem 4.2. The following holds:

(i) The true parameter $\theta_0$ is an interior point of $\Theta$.
(ii) The parameter space $\Theta$ is compact.
(iii) For all $u, w$, the mapping $\theta \mapsto M_{\theta}(u, w)$ is three times differentiable and its third order derivative is bounded uniformly in $u, w, \theta$.
(iv) The matrix $\nabla^2 L(\theta_0)$ is positive definite.
(v) The class $\{(t, z) \in \mathbb{R}_+ \times \mathcal{Z} \mapsto I(t \leq \phi_0(z, u)), \theta \in \Theta \}$ is Donsker for all $u \in \mathbb{R}_+, d \in \{0, 1\}$.
(vi) For all $u \in \mathbb{R}_+$, there exists a constant $C_u > 0$ such that $|\phi_0(z, u) - \phi_{0d}(z, u)| \leq C_u ||\theta - \theta_0||$ for all $\theta \in \Theta, d \in \{0, 1\}, z \in \mathcal{Z}$.
(vii) The density of $T$ given $Z, D$ is uniformly bounded.

These standard and mild conditions depend simultaneously on the regularity of the mapping $\theta \mapsto \phi_0$ and on the distribution of $(U, Z, W)$.

Theorem 4.1. Under Assumptions 2.1, 2.2, 2.3, 3.4, 4.1, and 4.2, there exists a $K \times K$ asymptotic variance matrix $\Sigma$ such that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

4.5. Bootstrap

Since the asymptotic variance matrix $\Sigma$ has a complicated expression (see the proof of Theorem 4.1), we rely on the nonparametric bootstrap for inference. Let $(Y_{bid}, \delta_{bid}, \bar{Z}_{bid}, \bar{D}_{bid}, W_{bid})_{i=1}^n$ be the bootstrap sample drawn with replacement from the original sample $\{X_i = (Y_i, \delta_i, \bar{Z}_i, \bar{D}_i, W_i)\}_{i=1}^n$. Let also $\hat{\theta}_b$ be the value of the estimator computed on the bootstrap sample $b$. The following result allows to build confidence intervals with the naive bootstrap using the empirical distribution of the estimates in the bootstrap samples.

Theorem 4.2. Under Theorem 4.1, we have $\sqrt{n}(\hat{\theta}_b - \hat{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma)^\star [P]$, where the convergence is for the law of $\hat{\theta}_b$ conditional on the original sample, in probability with respect to the original sample. In other words, $\sup_{t} P^\star(\sqrt{n}(\hat{\theta}_b - \hat{\theta}) \leq t) \leq P(\sqrt{n}(\hat{\theta} - \theta_0) \leq t) = o_P(1)$, where $P^\star$ stands for the probability law conditionally on the original data.

5. Numerical Experiments

5.1. Simulations

For a sample of size $n$, and for $i = 1, \ldots, n$, we generate the instrument $W_i$ and $U_i$ from two independent exponential distributions with parameters equal to one. We also generate an additional error term $r_i \sim \text{Exp}(1)$. The treatment time $Z_i$ is taken equal to $Z_i = \sqrt{2r_i} U_i I_i W_i^{\alpha}$, where the parameter $\alpha$ controls the level of endogeneity, and the parameter $\beta$ controls the strength of the instrument. When $\beta = 0$, the treatment time, $Z_i$, is independent of $U_i$, and when $\beta = 0$, the instrument cannot explain any of the variation in the treatment time. Finally, let $T_i = \phi(Z_i, U_i) = \phi_0(U_i)I(\phi_0(U_i) < Z_i) + \phi_1(Z_i, U_i)I(\phi_0(U_i) \geq Z_i)$.

For $\phi$, we consider the two parametric models of Section 4.1. For the Weibull model (example 1), we take the true parameter vector $\theta = (\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11})^\top = (1, 2, 1.5, 2)^\top$. For the log-normal model (example 2), the true parameter vector is $\theta = (\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11})^\top = (0, 1, 1, 1)^\top$. For both designs, we consider $\alpha \in [0.25, 0.75]$ to vary the level of endogeneity, and $\beta \in [0.5, 1]$ to vary the strength of the instrument. We further take censoring into account as follows:

(a) $T_i$ is not censored ($C_i = \infty$).
(b) In Setting 1 (Weibull) we take $C_i - 0.3 \sim \text{Exp}(2)$, and in Setting 2 (log-normal), $\log(C_i) \sim N(1, 1)$. The parameters of the distribution of $C_i$ are chosen in such a way that about 20% of the observations are censored.

The sample size is fixed to $n \in \{500, 1000, 3000\}$. We thus have a total of $3 \times 2^4 = 48$ simulation schemes, and we run $R = 1000$ replications for each scheme. The optimization algorithm is started at 100 random values. Each of these starting values yields a local minimum. The local minimum which leads to the lowest value of the objective function corresponds to the estimate. We use a grid $u_1, \ldots, u_{100}$ of 100 values of $U_i$, where $u_1$ (respectively, $u_{100}$) is the 0.025 (respectively, 0.975) quantile of the unit exponential distribution and the points are equally spaced.

We report the bias and standard error of the estimator and the coverage of the bootstrap percentile confidence intervals at level 90%, 95%, and 99%. As the computation time increases with the sample size, the coverage of the confidence intervals are evaluated using the Warp-Speed method of Giacomini et al. (2013), where only one bootstrap resampling is used for each simulated sample. We also provide the average number of treated units observed, $\bar{D}$, and the average number of uncensored observations, $\bar{r}$.

We summarize the simulations results for the Weibull design with censoring in Table 1. Between 33% and 43% of observations are treated. With a strong instrument ($\beta = 1$), the bias is low, even when $n = 500$. The coverage of the confidence intervals improves when $n$ grows and are relatively close to nominal for $n = 3000$. When the instrument is weaker ($\beta = 0.5$), the estimator exhibits good performance in terms of bias when $n = 1000$ or $n = 3000$. The results for the Weibull design without censoring and the log-normal model (with and without censoring) are reported in Section S.4 of the online appendix.

5.2. Empirical Application

We use data from a large Belgian insurance company. This insurer offers a product to other companies, which consists in paying for the salaries of their clients’ workers who are on medical leave because of burnout. This insurance product also contains the possibility of offering a free therapy for these workers suffering from burnout. The goal of the company is to reduce the duration of medical leave (our duration variable in this application) and, hence, it would like to know the effect of the timing of the therapy (our treatment) on the duration of medical leave.
The employees on medical leave are assigned to one of several partner institutions for medical care, which are responsible for the treatment time-varying. The treatment time is uncensored, the average is 112 days and the median is equal to 90 days.

We take advantage of the treatment assignment mechanism to evaluate the causal effect of the time-to-treatment. The employees on medical leave are assigned to one of several partner institutions for medical care, which are responsible for the treatment time-varying. The treatment time is uncensored, the average is 112 days and the median is equal to 90 days.

The exact date of the start of the therapy varies depending on the availability of the partner and the patient, which makes the treatment time-varying.

The assignment by the insurance company of the partner was only based on geographical distance, which should make it exogenous. Moreover, partners are more or less likely to offer a therapy to assigned patients, hence, the propensity of the partner to offer treatment has an impact on the time-to-treatment. We use the proportion of patients who are given the possibility to offer treatment has an impact on the time-to-treatment. The exact date of the start of the therapy varies depending on the availability of the partner and the patient, which makes the treatment time-varying.

The sample consists of 838 individuals who entered medical leave for burnout between 2017 and 2020. Their age at the start of the medical leave was between 30 and 39 years old. They are observed until one of the following events happens: their medical leave ends, their medical leave exceeds 2 years, or the study is ended (at the end of 2020). In the latter two cases the individual is censored. Since the duration of follow-up depends on external factors, we expect censoring to be uninformative. Around 41% of the observations are censored and 48% are treated before censoring. The average (respectively, median) duration of medical leave for uncensored observations is 189 days (respectively, 159 days). For the observations for which the treatment time is uncensored, the average is 112 days and the median is equal to 90 days.

We take advantage of the treatment assignment mechanism to evaluate the causal effect of the time-to-treatment. The employees on medical leave are assigned to one of several partner institutions for medical care, which are responsible for the treatment time-varying. The treatment time is uncensored, the average is 112 days and the median is equal to 90 days.

The exact date of the start of the therapy varies depending on the availability of the partner and the patient, which makes the treatment time-varying.

The assignment by the insurance company of the partner was only based on geographical distance, which should make it exogenous. Moreover, partners are more or less likely to offer a therapy to assigned patients, hence, the propensity of the partner to offer treatment has an impact on the time-to-treatment. We use the proportion of patients who are given the possibility to offer treatment has an impact on the time-to-treatment. The exact date of the start of the therapy varies depending on the availability of the partner and the patient, which makes the treatment time-varying.

The sample consists of 838 individuals who entered medical leave for burnout between 2017 and 2020. Their age at the start of the medical leave was between 30 and 39 years old. They are observed until one of the following events happens: their medical leave ends, their medical leave exceeds 2 years, or the study is ended (at the end of 2020). In the latter two cases the individual is censored. Since the duration of follow-up depends on external factors, we expect censoring to be uninformative. Around 41% of the observations are censored and 48% are treated before censoring. The average (respectively, median) duration of medical leave for uncensored observations is 189 days (respectively, 159 days). For the observations for which the treatment time is uncensored, the average is 112 days and the median is equal to 90 days.
For both the Weibull model and the log-normal model, we estimate the parameters using 100 random starting values around the values corresponding to the naive fit of the Weibull distribution or the log-normal distribution to the data of durations and censoring indicators. We use a grid \( u_1, \ldots, u_{100} \) of 100 values of \( U \), where \( u_1 \) is the 0.025 quantile of the unit exponential distribution and the points are equally spaced. The upper bound \( u_{100} \) is chosen such that
\[
\lim_{u \to 0} \frac{\varphi_{\hat{\theta}}(u, 0)}{\varphi_{\hat{\theta}}(u, 1)} < c_0 \quad \text{and} \quad \lim_{u \to 0} \frac{\varphi_{\hat{\theta}}(\tilde{Z}_i(u), u_{100})}{\varphi_{\hat{\theta}}(\tilde{Z}_i, u_{100})} < c_0, \quad i \in \{1, \ldots, n : \tilde{D}_i = 1\},
\]
where \( c_0 \) is the maximum follow-up time, is equal to 2 years. Condition (12) ensures that Assumption (I) (ii) is satisfied for some \( \Theta \) in a neighborhood of \( \hat{\Theta} \). Following this approach, we chose \( u_{100} \) equal to the 0.9 (respectively, 0.8) quantile of a unit exponential for the Weibull (respectively, log-normal) model. The curves of the hazard rates for the individuals who are never treated \((z = \infty)\) and those who received treatment at time \( 0 \) \((z = 0)\) corresponding to the estimated parameters are plotted in Figure 2 for the Weibull model and in Figure 3 for the log-normal model.

These hazard rates are computed by plug-in of the estimates in (8) and (9). By definition of our Weibull and log-normal models, for an arbitrary value of \( z \) the structural hazard rate at \( t \) under treatment at time \( z \) is equal to that of the never treated for \( t \leq z \) and is equal to that of the treated at time 0 for \( t > z \). Hence, in Figures 2 and 3, the estimated structural hazard under time-to-treatment equal to \( z \) corresponds to the red curve before \( z \) and then "jumps" to the blue dashed curve. We see that the therapy appears to increase the hazard rate for all possible treatment timings. As a result, the treatment should be administered at time 0 in order to minimize the duration of medical leave. Bootstrap confidence intervals for the hazard rates are given in Section S.5 of the online appendix. Note that the estimated hazards exhibit different shapes under the two models. This is to be expected since the models assume different parametric forms. However, the treatment significantly increases the hazard rate in both models (see the bootstrap confidence intervals in the online appendix). The robustness of this conclusion to the choice of the model constitutes statistical evidence supporting the efficacy of the therapy.

6. Concluding Remarks

This article develops an instrumental variable approach to estimate the causal effect of the time until a treatment is started on a possibly right-censored duration outcome. The time-to-treatment \( Z \) corresponds to the jump of a counting process with single jump. As an extension, it would be of interest to consider procedures where \( Z \) is a more general process. Another possible research direction could be to derive semiparametric identification results for the model. We conjecture that these results could be achieved under weaker assumptions than the completeness conditions of Section 3.

Supplementary Materials

The online appendix contains the proof of the results, additional interesting results and simulations.

Acknowledgments

The authors thank an Editor, an Associate Editor and two referees of the Journal of Business & Economic Statistics for helpful comments which greatly improved the article.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Financial support from the European Research Council (2016-2021, Horizon 2020 / ERC-grant agreement No. 694409) is gratefully acknowledged by Jad Beyhum and Ingrid Van Keilegom. Samuelle Centorrino would like to thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by an NSF grant (#1531492). Jean-Pierre Florens acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d’Avenir, grant ANR-17-EURE-0010).

References

Abbring, J. H., and Van den Berg, G. J. (2003), “The Nonparametric Identification of Treatment Effects in Duration Models,” Econometrica, 71, 1491–1517. [733,734]
Andrews, D. W. K. (2017), “Examples of \( L^2 \)-Complete and Boundedly-Complete Distributions,” Journal of Econometrics, 199, 213–220. [736]
Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), “Identification of Causal Effects Using Instrumental Variables,” Journal of the American Statistical Association, 91, 444–455. [733]
Beyhum, J., Florens, J.-P., and Keilegom, I. V. (2022), “Nonparametric Instrumental Regression with Right Censored Duration Outcomes,” Journal of Business & Economic Statistics, 40, 1034–1045. [733]
Blanco, G., Chen, X., Flores, C. A., and Flores-Lagunes, A. (2020), “Bounds on Average and Quantile Treatment Effects on Duration Outcomes Under Censoring, Selection, and Noncompliance,” Journal of Business & Economic Statistics, 38, 901–920. [733]
Brookhart, M. A., Wang, P., Solomon, D. H., and Schneeweiss, S. (2006), “Evaluating Short-Term Drug Effects Using a Physician-Specific Prescribing Preference as an Instrumental Variable,” Epidemiology (Cambridge, Mass.), 17, 268. [740]
Brown, D. J., and Wegkamp, M. H. (2002), “Weighted Minimum Mean-Square Distance from Independence Estimation,” Econometrica, 70, 2035–2051. [738]
Canay, I. A., Santos, A., and Shaikh, A. M. (2013), “On the Testability of Identification in Some Nonparametric Models with Endogeneity,” Econometrica, 81, 2535–2559. [737]
Cazals, C., Fève, F., Flores, J.-P., and Simar, L. (2016), “Nonparametric Instrumental Variables Estimation for Efficiency Frontier,” Journal of Econometrics, 190, 349–359. [736]
Chen, X., Chernozhukov, V., Lee, S., and Newey, W. K. (2014), “Local Identification of Nonparametric and Semiparametric Models,” Econometrica, 82, 785–809. [736]
Chen, Y., and Briesacher, B. (2011), “Use of Instrumental Variable in Prescription Drug Research with Observational Data: A Systematic Review,” Journal of Clinical Epidemiology, 64, 687–700. [740]
Chernozhukov, V., Fernández-Val, I., and Kowalski, A. E. (2015), “Quantile Regression with Censoring and Endogeneity,” Journal of Econometrics, 186, 201–221. [733]
Chernozhukov, V., and Hansen, C. (2005), “An IV Model of Quantile Treatment Effects,” Econometrica, 73, 245–261. [732,734,735,736,737]
Cui, Y., Michael, H., Tanser, F., and Tchetgen Tchetgen, E. (2023), “Instrumental Variable Estimation of the Marginal Structural Cox Model for Time-Varying Treatments,” Biometrika, 110, 101–118. [733]
Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011), “Nonparametric Instrumental Regression,” Econometrica, 79, 1541–1565. [736]
D’Haultfoeuille, X. (2011), “On the Completeness Condition in Nonparametric Instrumental Problems,” Econometric Theory, 27, 460–471. [736]
Frandsen, B. R. (2015), “Treatment Effects with Censoring and Endogeneity,” Journal of the American Statistical Association, 110, 1745–1752. [733]
Giacomini, R., Politis, D. N., and White, H. (2013), “A Warp-Speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators,” Econometric Theory, 29, 567–589. [739]
Heckman, J. J., and Navarro, S. (2007), “Dynamic Discrete Choice and Dynamic Treatment Effects,” Journal of Econometrics, 136, 341–396. [733]
Hernán, M. A., and Robins, J. M. (2010), Causal Inference: What If, Boca Raton, FL: Chapman & Hall/CRC. [733]
Hu, Y., and Shiu, J.-L. (2018), “Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness,” Econometric Theory, 34, 659–693. [736]
Kastoryano, S., and van der Klaauw, B. (2022), “Dynamic Evaluation of Job Search Assistance,” Journal of Applied Econometrics, 37, 227–241. [733]
Lechner, M. (2009), “Sequential Causal Models for the Evaluation of Labor Market Programs,” Journal of Business & Economic Statistics, 27, 71–83. [733]
Lechner, M., and Miquel, R. (2010), “Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions,” Empirical Economics, 39, 111–137. [733]
Michael, H., Cui, Y., Lorch, S. A., and Tchetgen Tchetgen, E. J. (2023), “Instrumental Variable Estimation of Marginal Structural Mean Models for Time-Varying Treatment,” Journal of the American Statistical Association, 1–12. [733]
Newey, W. K., and Powell, J. L. (2003), “Instrumental Variable Estimation of Nonparametric Models,” Econometrica, 71, 1565–1578. [736]
Richardson, A., Hudgens, M. G., Fine, J. P., and Brookhart, M. A. (2017), “Nonparametric Binary Instrumental Variable Analysis of Competing Risks Data,” Biostatistics, 18, 48–61. [733]
Sant’Anna, P. H. (2016), “Program Evaluation with Right-Censored Data,” arXiv preprint arXiv:1604.02642. [733]
----- (2021), “Nonparametric Tests for Treatment Effect Heterogeneity with Duration Outcomes,” Journal of Business & Economic Statistics, 39, 816–832. [733]
Sianesi, B. (2004), “An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s,” Review of Economics and Statistics, 86, 133–155. [733]
Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015), “Instrumental Variable Estimation in a Survival Context,” Epidemiology (Cambridge, Mass.), 26, 402. [733]
Van den Berg, G. J., Boven, P., and Mammen, E. (2020), “Nonparametric Instrumental Variable Methods for Dynamic Treatment Evaluation,” Review of Economics and Statistics, 102, 355–367. [733]
Van den Berg, G. J., Bozio, A., and Costa Dias, M. (2020), “Policy Discontinuity and Duration Outcomes,” Quantitative Economics, 11, 871–916. [733]
Vikström, J. (2017), “Dynamic Treatment Assignment and evaluation of Active Labor Market Policies,” Labour Economics, 49, 42–54. [733]