


# BMJ Open Do studies evaluating early-life policy interventions fully adhere to the critical conditions of difference-in-differences? A systematic review

Anouk Klootwijk <sup>1,2</sup>, Jeroen Struijs,<sup>1,2</sup> Annelieke Petrus,<sup>2</sup> Marlin Leemhuis,<sup>1</sup> Mattijs Numans,<sup>2</sup> Eline de Vries<sup>3</sup>

**To cite:** Klootwijk A, Struijs J, Petrus A, *et al.* Do studies evaluating early-life policy interventions fully adhere to the critical conditions of difference-in-differences? A systematic review. *BMJ Open* 2024;**14**:e083927. doi:10.1136/bmjopen-2024-083927

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2024-083927>).

Received 03 January 2024  
Accepted 03 May 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

<sup>1</sup>Department for Population Health and Health Services Research, National Institute for Public Health and the Environment, Bilthoven, Netherlands

<sup>2</sup>Health Campus The Hague/ Department of Public Health and Primary Care, Leiden University Medical Center, The Hague, Netherlands

<sup>3</sup>Department for Health Economics and Health Services Research, National Institute for Public Health and the Environment, Bilthoven, Netherlands

## Correspondence to

Anouk Klootwijk;  
[anouk.klootwijk@rivm.nl](mailto:anouk.klootwijk@rivm.nl)

## ABSTRACT

**Objectives** To assess the reporting and methodological quality of early-life policy intervention papers that applied difference-in-differences (DiD) analysis.

**Study design** Systematic review.

**Data sources** Papers applying DiD of early-life policy interventions in high-income countries as identified by searching Medline, Embase and Scopus databases up to December, 2022.

### Study eligibility criteria, participants and interventions

Studies evaluating policy interventions targeting expectant mothers, infants or children up to two years old and conducted in high income countries were included. We focused on seven critical conditions of DiD as proposed in a comprehensive checklist: data requirements, parallel trends, no-anticipation, standard statistical assumptions, common shocks, group composition and spillover.

**Results** The DiD included studies (n=19) evaluating early-life policy interventions in childhood development (n=4), healthcare utilisation and providers (n=4), nutrition programmes (n=3) and economic policies such as prenatal care expansion (n=8). Although none of the included studies met all critical conditions, the most reported and adhered to critical conditions were data requirements (n=18), standard statistical assumptions (n=11) and the parallel trends assumption (n=9). No-anticipation and spillover were explicitly reported and adhered to in two studies and one study, respectively.

**Conclusions** This review highlights current deficiencies in the reporting and methodological quality of studies using DiD to evaluate early-life policy interventions. As the validity of study conclusions and consequent implications for policy depend on the extent to which critical conditions are met, this shortcoming is concerning. We recommend that researchers use the described checklist to improve the transparency and validity of their evaluations. The checklist should be further refined by adding order of importance or knock-out criteria and may also help facilitate uniform terminology. This will hopefully encourage reliable DiD evaluations and thus contribute to better policies relating to expectant mothers, infants and children.

## INTRODUCTION

Difference-in-differences (DiD) is a widely applied quasi-experimental study design

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ A comprehensive literature search was undertaken across major existing databases.
- ⇒ A formal checklist was used to critically appraise studies that applied difference-in-differences (DiD) to evaluate early-life policy interventions on their adherence to seven critical conditions underlying DiD.
- ⇒ The checklist we used was not exhaustive, and inclusion of other critical conditions and adding an order of importance of critical conditions might have resulted in slightly different conclusions.

that is commonly used to evaluate public policy across diverse research areas such as economics and health services research.<sup>1-5</sup> Over the last decade, studies using DiD have also emerged specifically in the evaluation of early-life policy interventions such as the Revised Special Supplemental Nutrition Programme for Women, Infants and Children and the State Children's Health Insurance Programme (SCHIP).<sup>6,7</sup> Early-life policy interventions generally promote healthy development, well-being and learning opportunities during the critical early years of a child's life.

Children's early experiences and environment have important long-term consequences in terms of health and social outcomes.<sup>8</sup> Early-life years, from conception up to 2 years of age, therefore offer a critical window of opportunity to shape the trajectory of a child's development and later life. When children miss out on the best potential start in life, cycles of poverty and disadvantage may recur for generations. Targets set by the 2030 Agenda for Sustainable Development, which have not been met even in high-income countries, involve issues such as poverty, exclusion and pollution, and thus threaten the mental well-being, physical health and



opportunities to develop skills within high-income countries.<sup>9</sup> Supporting a good start in life by providing early-life policy interventions to all young children and families is considered one of the most powerful and cost-effective equalisers, ensuring that even the most vulnerable children reach their full potential.<sup>8</sup>

To optimally inform policymakers on the effects of policy interventions in early-life years, reliable evidence obtained with robust evaluations is needed.<sup>10</sup> These evaluations can provide information on access, participation, equality and the quality of provision. While the strongest evaluation designs compare children and parents who receive programme services with a comparison group of children and parents who do not receive those services, sometimes this is not possible for practical or ethical reasons. In settings where randomised controlled trials are not feasible or unethical, other evaluation designs are used to study causal effects. These include quasi-experimental designs, of which DiD is frequently used in public health research.<sup>1</sup>

DiD gained popularity as a result of the intuitive conceptual design, coupled with the increasing availability of longitudinal data. The underlying principle of a DiD design involves (1) the availability of an intervention, (2) a preintervention period, (3) a postintervention period, (4) a treatment group and (5) a control group.<sup>11</sup> In a DiD, the difference in outcomes between the treatment and control group at baseline (difference 1) and subsequently during the postintervention period (difference 2) represents the estimated average treatment effect of the treated (difference 2 minus difference 1),<sup>4,11</sup> hence explaining the name ‘difference-in-differences’. By combining the pre–post comparisons with across treatment and control group comparisons, confounding due to both (unobserved) time and selection bias can be excluded.<sup>13</sup>

Despite the apparent simplicity and intuitive nature of the DiD concept, several challenges hinder the drawing of valid conclusions. There is a vast body of literature showing that specification choices, such as the choice of a control group, can impact point estimates and statistical significance.<sup>3,4,12–14</sup> These specification choices in turn affect the validity of policy recommendations based on the estimates. In order to be able to draw valid causal conclusions, assumptions and other conditions of DiD, such as data requirements, parallel trends assumption and no-anticipation, need to be met, hereafter collectively referred to as ‘critical conditions’. Ryan *et al* proposed a checklist encompassing seven critical DiD conditions, including checks and mitigation strategies for applied research.<sup>11</sup> While recognising that the checklist by Ryan *et al* is not all-encompassing, it is the most comprehensive formalised checklist currently available for DiD critical conditions.

To date, the extent to which studies evaluating early-life policy interventions meet these ‘critical conditions’ is unknown. As the validity of causal conclusions and consequent policy decisions strongly depend on the extent to

which these critical conditions are met, better insight into this issue is crucial. Therefore, the aim of this study was to systematically analyse the scientific literature to identify and assess the reporting and methodological quality of studies evaluating early-life policy interventions using DiD.

## MATERIALS AND METHODS

### Eligibility criteria

The PRISMA (Preferred Reporting Items for Systematic Reviews) 2020 statement was used to ensure the validity and reliability of the selection procedure.<sup>15</sup> We included studies based on the following three eligibility criteria. First, studies focusing on early-life policies or interventions were included, defined as regulations, legislations, fiscal policies and mandates targeting expectant mothers, infants and children during the first two years of life, hereafter collectively referred to as ‘early-life policy interventions’. Second, studies were required to use DiD as their main analysis technique. Third, studies were included if they studied populations from high-income countries as defined by the World Bank.<sup>16</sup> We excluded letters, commentaries, theoretical simulations, studies not written in English and those not published in peer-reviewed journals.

### Search strategy

Our search strategy was set up with the help of a librarian (available in online supplemental data S1). Search terms were based on the eligibility criteria and described early life and DiD. Searches were conducted within Medline, Embase and Scopus, along with Internet searches via Google and reference tracking, encompassing data available until December 2022. All search results were stored in EndNote.

### Selection procedure

After duplicate removal, two researchers (AK and ML) independently performed screening of titles and abstracts to assess eligibility. Inconsistencies between researchers were resolved through discussion and a third researcher (JS) acted as mediator when necessary.

### Data extraction

We extracted data from the included studies using two predefined extraction tables. The first table consisted of general information: authors, publication year, the evaluated policy intervention (including name, objective and implementation period), the level/setting on which the policy intervention was implemented, data sources, study population and main outcomes. The second table concerned whether critical conditions of DiD, as proposed in the checklist of Ryan *et al*, were explicitly reported and if these critical conditions were met.<sup>11</sup> AK extracted the data, while ML randomly verified 25% of the studies for consistency. Given the minimal inconsistencies

encountered, we did not expand further data extraction checking.

### Critical appraisal

The included studies were critically appraised in relation to the aforementioned seven critical conditions of DiD. Specifically, we assessed whether the criteria were reported and if they were met as proposed in the checklist of Ryan *et al.*<sup>11</sup> We appraised the seven critical conditions based on the information reported in the methods section within each study. If information was not present in the methods section, we examined the results section, discussion section and appendices for useful information. Concerning reporting quality, per critical condition we assessed whether a condition was explicitly mentioned ('yes') or omitted ('no'). Per critical condition, we noted '+' if the study indicated criteria fulfilment, '-' if the study did not meet criteria and '?' indicating unclear or missing information. The critical appraisal was prepared and executed by AK and thoroughly discussed with EdV and JS. After consensus was reached on the critical appraisal, a statistician was consulted for a final verification.

### Critical conditions of DiD based on checklist by Ryan *et al.*<sup>11</sup>

Ryan *et al* described seven critical conditions for DiD: (1) data requirements, (2) parallel trend, (3) no-anticipation, (4) standard statistical assumptions, (5) common shocks, (6) group composition and (7) spillover. An overview of the assessment per critical condition, and what to do if the critical condition was violated is presented in [table 1](#).

### Data requirements

To implement a DiD design, longitudinal data must be available on study outcomes for both the treatment and control groups, including at least one time period both

before and after implementation of the intervention. This critical condition is directly observable and if violated no valid inferences can be made using DiD according to the checklist by Ryan *et al.*<sup>11</sup> Studies using data from the preintervention and postintervention period for both the treatment and control group were appraised as '+' for data requirements.

### Parallel trends assumption

Parallel trends is considered a key assumption for DiD.<sup>1 3 4 12 17-19</sup> The parallel trends assumption requires the treatment and control groups to change at the same rate prior to the intervention, although treatment and control groups may show different levels of the outcome in question. The parallel trends assumption is needed to calculate the counterfactual, defined as the likely outcome for the treatment group in the postintervention period had the treatment group not been exposed to the intervention. If the parallel trends assumption holds preintervention, it is possible to estimate the counterfactual postintervention based on the control group's preintervention trend. This parallel trends assumption is crucial if we wish to validly attribute a difference between the differences in outcomes of groups to the policy intervention, rather than to pre-existing differential trends in outcomes and, thereby exclude selection bias. Studies were viewed as adhering to the parallel trends assumption if they assessed whether linear preintervention trends differed statistically between the treatment and control groups (appraised as '+'). This was appraised by testing the significance of the interaction term between the time trend and the treatment group at multiple data points in the preintervention period. In addition to the definition of Ryan *et al* as regards the parallel trends assessment,<sup>11</sup>

**Table 1** Critical conditions of DiD and mitigations if violated based on checklist by Ryan *et al.*<sup>11</sup>

Critical condition	Assessment	Mitigation strategies if violated
Data requirements	Directly observable	NA
Parallel trends assumption	Assess whether preintervention trends are parallel between treatment and control group (placebo test plus event study)	If multiple control groups are available, match treatment and control units
No-anticipation	Assess whether baseline outcome is correlated with the probability of assignment to the treatment across the study period for both treatment and control group	If multiple control groups are available, match treatment and control units
Standard statistical assumptions	Assess violations of SEs	Clustered SEs and permutation tests are recommended
Common shocks	Generally not testable, but other factors than the intervention that may affect outcomes for either the treatment or control group should be taken into account in the interpretation	NA
Group composition	Assess the difference in observed covariates in both the preintervention and postintervention period between the treatment and control group and test differential drop-out rates between treatment and control group	Control the analysis for covariates with observed differences between treatment and comparison group before and after the intervention
Spillover	Assess whether the control group shows deviation from existing trend in the outcome concurrent with the intervention	Choose an alternative control group that is not subject to spillovers

DiD, difference-in-differences; NA, not applicable.



we also judged placebo tests at multiple data points in the preintervention period and event studies as adhering to a parallel trends assessment. A placebo test is a technique that shifts the time of intervention to before the actual time of intervention, at which point DiD is expected to yield a no-effect estimate.<sup>20</sup> If a placebo test measures an effect, the parallel trends assumption must be rejected. An alternative method to test for parallel trends, the event study graph, shows differences between the treatment and control groups at different time points before and after the intervention.<sup>21</sup> Event studies depict whether differences exist between the intervention and control groups at different time points before and after the intervention. Both placebo tests and event studies were appraised as ‘+’ for parallel trends assessment.

### No-anticipation

Another frequently mentioned critical condition of DiD is no-anticipation.<sup>14 22 23</sup> No-anticipation states that outcome levels in the preintervention period should not be associated with the probability of assignment to the intervention. This critical condition is violated if treatment effects are disproportionately present in the treatment group in preintervention periods, which may lead to misinterpretation of the treatment effect.<sup>24</sup> Violation of no-anticipation is plausible in many setups, especially if an intervention is announced in advance, potentially leading to behavioural change in response to information about a policy rather than the policy itself. We appraised studies as adhering to this critical condition if they assessed whether the outcome in the preintervention period correlated with change in performance of the treatment group.<sup>11</sup> As with the parallel trends assumption, matching treatment and control group for preintervention levels is recommended to reduce bias.<sup>11</sup> It is important to assess no-anticipation, even if trends between treatment and control groups are parallel in the preintervention period.

### Standard statistical assumptions

DiD is performed using regression analysis and is therefore subject to standard statistical assumptions. Point estimates of policy intervention effects can be easily generated by calculating the difference in means for the outcome between treatment and control groups, before and after the intervention was implemented. Regression models enable calculation of DiD estimates, with CIs for variance. Furthermore, regression analysis allows more advanced specifications to be developed, improving the accuracy of point estimates and statistical inference. However, it is critical that violations of standard statistical assumptions are addressed appropriately. Ryan *et al* recommend the use of clustered SEs to account for heteroskedasticity at the cluster level, as this results in lower false rejection rates.<sup>11</sup> Ryan *et al* also recommend performance of permutation tests used for exact inference if ‘assumptions underlying other variance estimators may be violated’.<sup>11</sup> There are no general recommendations for the level of clustering, as this depends on the sampling (i.e., the level

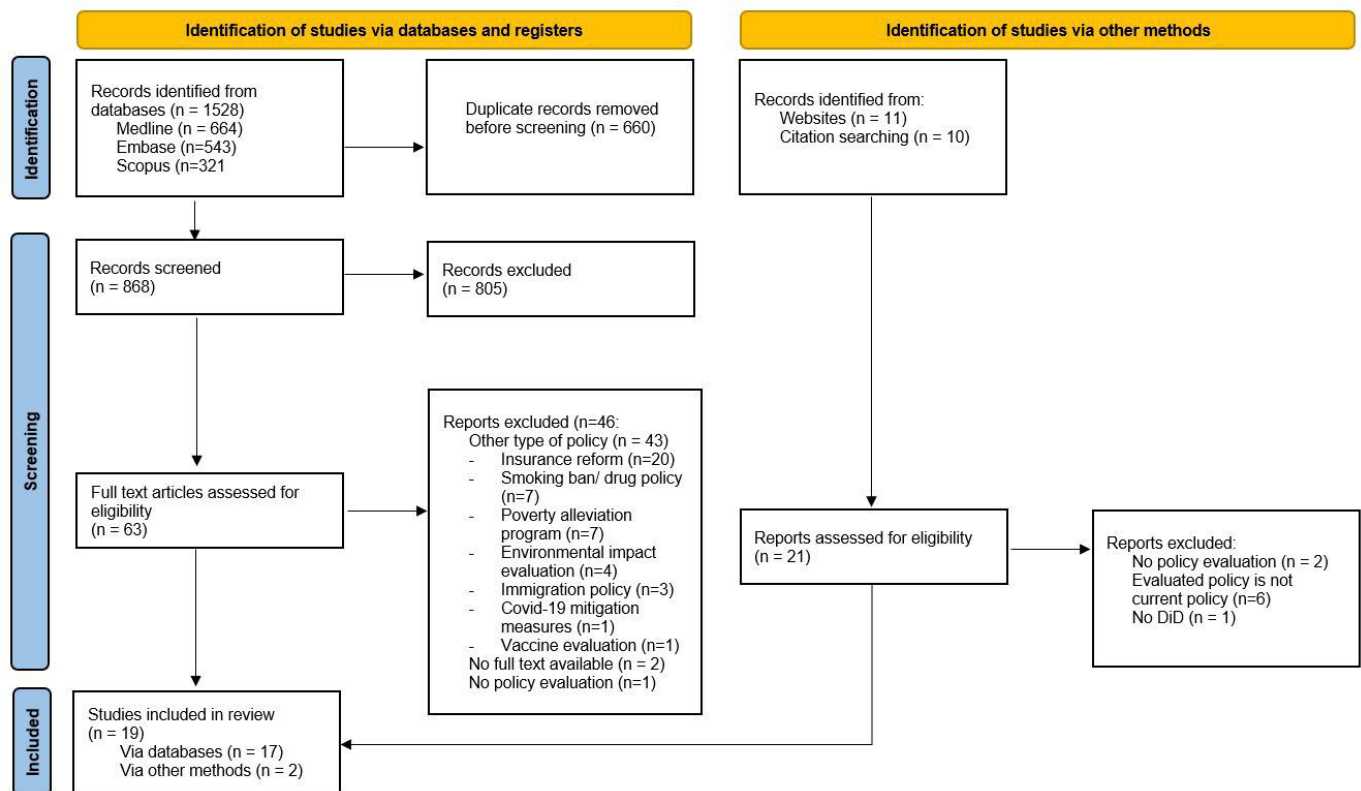
of clustering depends on how the sample is drawn from the population) and the parameter of interest (the level of clustering depends on the treatment assignment level). We appraised studies as adhering to standard statistical assumptions if they applied clustered SEs or permutation tests.

### Common shocks

‘Common shocks’ refers to other phenomena occurring at the same time or after the start of treatment which equally affect the treatment and control groups. This critical condition is generally not testable, but factors other than the intervention that may affect outcomes for either the treatment or control group should be taken into account. Changes in unobserved factors over the study period, for example, self-selection into treatment which reflect an increased but unobserved interest in improving the outcome, may result in ‘expected gains bias’ if the changing unobserved factor has led to effects not attributable to the programme itself, resulting in an overestimation of the estimated programme effect. This assumption is not directly testable and if violated no valid inferences can be made using DiD according to Ryan *et al*.<sup>11</sup> Ryan *et al* did not propose a test for this assumption, although several strategies exist to test for common shocks. We viewed additional analyses of common shocks, for example, inclusion of fixed effects or addition of extra control groups, as adhering to this critical condition.

### Group composition

DiD also rests on the assumption that the composition of both the treatment and control group remains constant over the course of the study period, including all unobserved factors affecting outcomes. However, compositional change can occur due to differential drop-out between treatment and control groups. To assess this critical condition, Ryan *et al* proposed testing for any difference in observed covariates between treatment and control groups before and after the intervention.<sup>11</sup> If differences are identified, they should be addressed by controlling for them in the analysis. DiD can be applied on two levels, the ‘group level’ and the ‘individual level’. In the group-level specification, data exist at the level at which the treatment occurs, for instance hospitals or neighbourhoods. Variation over time in the composition of confounders between groups can confound effect estimates. To compensate for this variation over time, group-level outcomes are often adjusted prior to estimation, thereby mitigating the effects of compositional change. On the individual level, compositional differences are taken into account by controlling for individual heterogeneity. Studies that assessed the difference in observed covariates between treatment and control groups before and after an intervention were appraised as adhering to the critical condition of group composition.



**Figure 1** Study selection flow diagram. DiD, difference-in-differences.

### Spillover

Spillover occurs if the control group is affected by the intervention, indicating an invalid DiD design. This critical condition can be assessed by testing whether the control group experiences deviation from an existing trend concurrent with the intervention. Studies that conducted this analysis were considered to have adhered to the critical condition of spillover. If there is no change in outcomes in control groups during the period of the intervention, this suggests no associated spillover effects. Spillover effects may be important when policy in one area affects neighbouring areas. In the event of spillover, choosing an alternative control group not subject to spillover is recommended if multiple control groups are available.<sup>11</sup> On a more aggregated level, for example switching the unit of analysis from patients to hospitals might be a solution, but it also changes the type of question that one can answer with the analysis. Spillover can be excluded through use of the Stable Unit of Treatment Value Assumption (SUTVA), used by Rubin,<sup>25</sup> Imbens<sup>26</sup> and Lechner,<sup>12</sup> an assumption in which units are unaffected by treatment for other units.<sup>14</sup>

### Patient and public involvement

Patients and the public are not involved in this study.

### RESULTS

The initial search identified 1528 studies (figure 1). After removal of duplicates (n=660), the titles and abstracts of 868

records were screened, resulting in 63 studies eligible for full text assessment. Reference tracking and internet searches identified 21 additional studies. On review of full-text records of 84 studies, 65 studies were excluded. The main reason for exclusion was evaluation of general types of policy interventions (n=43), for example, insurance reforms, rather than policy interventions specifically targeted at early life. Ultimately, 19 papers fulfilled our criteria and could be included in this systematic review (figure 1). All papers were published after 2010, and 14 studies were published just within the last five years. US studies predominated (n=12), followed by European studies (n=6) and a single Chilean study (n=1), with included studies covering the following areas: early childhood development (n=4), healthcare utilisation and providers (n=4), nutrition programmes (n=3) and various economic policies such as prenatal care expansion and family leave (n=8) (online supplemental material S2, table 1). Some studies covered similar initiatives: the State Children's Health Insurance Programme in the USA (n=2),<sup>7 27</sup> Revised Special Supplemental Nutrition Programme for Women, Infants and Children, also in the USA (n=3),<sup>6 28 29</sup> paid family leave (n=2)<sup>30 31</sup> and the Salut Programme (n=2).<sup>32 33</sup> Eight early-life policy interventions were implemented at the state level in the USA.

### Critical conditions of DiD studies based on the checklist by Ryan *et al*<sup>1</sup>

Below we discuss whether included studies reported and met each of the seven critical conditions.

### Data requirements

All included studies used data on both the preintervention and postintervention periods for the treatment and control groups, with the exception of a study by Cygan-Rehm and Karbownik.<sup>34</sup> Data on a separate control group were lacking in this study, as the treatment group served as its own control group for the corresponding weeks in a distinct year prior to the intervention.

### Parallel trends assumption

Of the 19 included studies, 11 explicitly addressed parallel or preintervention trends.<sup>6 19 28–31 34–38</sup> Of these, the methodology of parallel trends assessment was in line with the proposed checklist method in nine studies. This involved regressing outcomes on the interaction term for the intervention and preintervention time trends, and observing the effect of interaction term significance<sup>6 29 30 35 37 38</sup> or by conducting an event study.<sup>31 34 36</sup> Six other studies reported alternatives to assess parallel trends that did not meet the criteria as proposed in the checklist. Reported alternative methods included visual inspection of preintervention trends only,<sup>28</sup> reporting preintervention differences in characteristics between the treatment and control group,<sup>7</sup> using an arbitrary intervention date,<sup>39 40</sup> conducting a placebo test at only one preintervention time point<sup>19</sup> and conducting a placebo test in which the outcome was regressed in lagged months.<sup>41</sup> These methods were appraised as ‘-’. The remaining four studies neither reported nor assessed parallel trends.

### No-anticipation

The critical condition of no-anticipation was reported and followed in two studies.<sup>34 36</sup> Cygan-Rehm and Karbownik investigated this interaction term between the outcome and the preintervention period when the policy was announced but not yet implemented.<sup>34</sup> Meinhofer *et al* inspected no-anticipation visually by depicted leads and lags alongside parallel trends, thus confirming to the checklist.<sup>36</sup> The remaining studies did not explicitly describe or assess no-anticipation.

### Standard statistical assumptions

Clustered SEs were reported and applied in 11 studies.<sup>27–29 31 34 36–38 40–42</sup> The clustering level was various at US state level,<sup>27 29 36 38 40</sup> county,<sup>42</sup> municipality,<sup>41</sup> local authority<sup>37</sup> or individual level.<sup>28 34</sup> One study used SEs with two lags.<sup>31</sup> The remaining eight studies did not apply clustered SEs. None of the included studies used permutation tests.

### Common shocks

Ten studies discussed the potential influence of other policy changes on the estimated effects.<sup>6 7 27 29–31 34 36 37 39</sup> Six of these studies performed an additional analysis to correct for common shocks,<sup>6 7 29 34 36 37</sup> such as a robustness check that tested the findings against secondary control variables.<sup>36</sup> The remaining nine studies neither reported nor conducted any additional analysis to correct for common shocks.

### Group composition

The composition of the treatment and control groups before and after the intervention was explicitly reported in eight studies.<sup>6 19 27–29 34 38 42</sup> Of these, six studies included testing of the difference in observed covariates between the treatment and control groups in both preintervention and postintervention periods, as proposed in the checklist (appraised as ‘+’). Eight studies controlled for changes in composition without initially checking compositional change,<sup>7 27 30 35 36 39 41 42</sup> and are thus appraised as ‘-’. The remaining five studies neither reported nor assessed composition of treatment and control group. Differential drop-out was not mentioned in any study.

### Spillover

Two studies explicitly mentioned potential spillover effects on older siblings, which may have impacted estimated effects towards zero.<sup>31 37</sup> Neither of these studies assessed whether the control group was affected by the intervention as proposed in the checklist. Alternatively, Cattan *et al* conducted an analysis excluding mothers with more than one child to determine whether this influenced their estimated effects.<sup>37</sup> Pihl and Basso did not assess potential spillover effects (both studies were appraised as ‘-’).<sup>31</sup> One study that only implicitly discussed spillover did assess this critical condition (and was appraised as ‘+’).<sup>34</sup> A residential mobility scenario was discussed in four studies, in which subjects might initially be exposed to an intervention in the original intervention area, but on relocation to the control area might be inadvertently included in the control group as ‘never treated’.<sup>32 37–39</sup> However, these four studies did not explicitly report spillover or conduct an assessment for this critical conditions (appraised as ‘-’). The remaining 13 studies neither reported nor assessed spillover effects.

Overall, studies varied markedly in the number and reporting of critical conditions (table 2), with no study reporting and adhering to all seven critical conditions. Three studies carried out appropriate assessment of either five<sup>29 36</sup> or six critical conditions,<sup>34</sup> all of which were published in 2022. Eight studies that adhered to two or less DiD critical conditions<sup>7 19 32 33 35 39 40 42</sup> were published between 2014 and 2021. Details of how authors of the included studies assessed and reported on critical conditions are presented in online supplemental material S3, table S2.

## DISCUSSION

This systematic review appraised the reporting and methodological quality of DiD studies evaluating early-life policy interventions. Specifically, we assessed whether authors considered the seven critical conditions of DiD and if the studies met these critical conditions as proposed in the checklist by Ryan *et al*.<sup>11</sup> Among the 19 included DiD studies, we found wide variation in both reporting quality and the number of critical conditions assessed, ranging from only one up to six per paper. The parallel trends

**Table 2** Reporting and methodological quality based on the critical conditions for DiD studies derived from Ryan *et al*<sup>11</sup>

Citation	Study (year)	Data requirements			Parallel trend			No-anticipation			Standard statistical assumptions			Common shocks			Group composition			Spillover		
		Reported	Criteria met	on	Reported	Criteria met	on	Reported	Criteria met	on	Reported	Criteria met	on	Reported	Criteria met	on	Reported	Criteria met	on	Reported	Criteria met	on
		on	met	on	on	met	on	on	met	on	on	met	on	on	met	on	on	met	on	on	met	on
19	Alcaide and Arranz (2021)	Y	+	Y	-	N	?	N	?	N	?	N	?	Y	+	N	?	Y	+	N	?	
37	Cattan <i>et al</i> (2021)	Y	+	Y	+	N	?	Y	+	Y	+	Y	+	N	+	N	?	N	+	Y	-	
41	Clarke <i>et al</i> (2020)	Y	+	N	?	N	?	Y	+	N	?	N	?	N	?	N	?	N	-	N	?	
34	Cygan-Rehm and Karbownik (2022)	N	-	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	+	
40	Dahlen <i>et al</i> (2017)	Y	+	N	-	N	?	Y	+	N	?	N	?	N	?	N	?	N	?	N	?	
7	Drewry <i>et al</i> (2014)	Y	+	N	?	N	?	N	?	Y	?	N	?	Y	+	N	?	N	-	N	?	
6	Guan <i>et al</i> (2020)	Y	+	Y	+	N	?	N	?	Y	?	N	?	Y	+	Y	+	Y	+	N	?	
32	Häggström <i>et al</i> (2017)	Y	+	N	?	N	?	N	?	N	?	N	?	N	?	N	?	N	?	N	-	
28	Hamad <i>et al</i> (2019)	Y	+	Y	-	N	?	Y	+	N	?	Y	+	Y	+	N	?	Y	+	N	?	
39	Janevic <i>et al</i> (2018)	Y	+	N	?	N	?	N	?	N	?	N	?	Y	+	N	?	N	-	N	-	
35	Jonge de <i>et al</i> (2019)	Y	+	Y	+	N	?	N	?	N	?	N	?	Y	+	N	?	N	-	N	?	
36	Meinhofer <i>et al</i> (2022)	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	?	N	-	N	?	
30	Montoya <i>et al</i> (2020)	Y	+	Y	+	N	?	N	?	N	?	N	?	Y	+	N	?	N	-	N	?	
13	Pihl and Basso (2019)	Y	+	Y	+	N	?	Y	+	Y	+	Y	+	N	?	N	?	N	?	Y	-	
33	Puikki-Brännström <i>et al</i> (2020)	Y	+	N	?	N	?	N	?	N	?	N	?	N	?	N	?	N	?	N	?	
29	Pulvera <i>et al</i> (2022)	Y	+	Y	+	N	?	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	?	
38	Rossin <i>et al</i> (2011)	Y	+	Y	+	N	?	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	-	
42	Swartz <i>et al</i> (2017)	Y	+	N	?	N	?	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	?	
27	Wherry <i>et al</i> (2017)	Y	+	N	?	N	?	Y	+	Y	+	Y	+	Y	+	Y	+	Y	+	N	?	

Some studies did adhere to the critical condition as proposed in the checklist, while not explicitly reported. Pulvera *et al* did not explicitly discuss clustered SEs, but stated the formula with clustered SEs in the appendix.<sup>29</sup> Spillover was not explicitly reported by Cygan-Rehm and Karbownik.<sup>34</sup>

+, meets criteria as proposed in the checklist; ?, notable to score; -, does not meet criteria; N, no; Y, yes.



assumption was the most frequently reported (n=11) and a majority of studies (n=9) assessed this critical condition in accordance with the checklist. By contrast, other critical DiD conditions, such as no-anticipation and spillover, were reported and assessed as proposed in the checklist in only two studies and one study, respectively.

As many studies inadequately assessed essential DiD critical conditions, our findings underline the pitfalls involved when assessing reported outcomes of studies using DiD to evaluate early-life policy interventions. None of the included studies fully met all critical conditions of DiD. Reviews of other quasi-experimental designs have reported slightly better rates of fulfilment of all quality appraisal criteria, for example, 12%–16% for regression discontinuity designs. Nonetheless, reviews of studies using other quasi-experimental designs have also stressed the need for clarity concerning systematic reporting.<sup>43 44</sup> Our results reconfirm the finding that standardised terminology and additional guidelines are crucial to improve adherence to critical conditions of quasi-experimental designs.

Our results also suggest that researchers applying DiD may be unfamiliar with all of the critical conditions pertaining to DiD. Consequently, the reliability of DiD effect estimates may be suspect and could be impacted by other confounding factors such as differential pre-intervention trends, anticipation, etc.<sup>3</sup> This is concerning, particularly because evaluations of early-life policy interventions influence policy when deciding on the implementation and upscaling of such interventions. Consequently, while policymakers may assume that studies published in peer-reviewed journals adhere to state-of-the-art methodologies, in fact important decisions may be based on information garnered from studies that failed to fully adhere to the critical conditions of DiD. At the moment, it is unclear to what extent this had negative consequences as we do not know how conclusions based on DiD evaluations might have differed if studies had fully adhered to the critical conditions of DiD.

As demonstrated by our findings, the included studies showed substantial inconsistencies in terms of focus on critical conditions of DiD. This might be explained by the rapid and ongoing evolution of DiD methodology, which poses challenges for all researchers. However, a recently published synthesis on DiD advances is available that offers concrete recommendations.<sup>14</sup> These advances can be broadly classified as modification of the DiD model with two time periods, a treatment group and a control group, which relax some critical conditions of the simplified DiD model such as variation in treatment timing.

This study had a number of strengths and limitations. To the best of our knowledge, this is the first review to assess adherence to critical conditions in DiD studies in the field of early-life policies. Together with our comprehensive search for papers evaluating early-life policy interventions with DiD, the checklist described by Ryan *et al* was particularly useful for critical appraisal of the DiD analysis, as this checklist is the most comprehensive available

in the current literature. A synthesis on DiD by Roth *et al* also included an alternative checklist that covered more in-depth checks for specific conditions,<sup>14</sup> for example, parallel trends and treatment timing. However, we felt that this checklist was limited in scope compared with the checklist by Ryan *et al*. Nevertheless, the checklist by Ryan *et al* has certain caveats.<sup>11</sup> The number of DiD critical conditions in the checklist was not exhaustive and inclusion of other critical conditions, for example SUTVA instead of spillover, might have resulted in different conclusions. Despite these concerns, we believe that key results of this study would not have changed had the checklist focused on related but different critical conditions. As the checklist by Ryan *et al* does not apply any form of ranking to critical conditions, another limitation may have been our equal appraisal of all critical conditions,<sup>11</sup> potentially resulting in undervaluation or overvaluation if certain critical conditions are actually more relevant than others. We also excluded studies not published in English, which may have omitted some relevant studies from the results. Nonetheless, our findings align with other systematic reviews that assessed adherence to the critical conditions of quasi-experimental designs, and we therefore consider our findings generalisable and relevant for researchers in countries other than those covered in our review.

### Future research

Integrating the critical conditions of DiD in a framework that acknowledges the key sources of bias for each critical condition will contribute to better causal inference, analogous to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) framework for observational study designs.<sup>15</sup> We recommend use of the checklist by Ryan *et al* as a starting point from which to develop a widely supported framework that will promote critical appraisal of DiD studies, help guide researchers and increase the transparency, replicability and credibility of policy-relevant evaluations.<sup>11</sup> The resulting checklist could be further developed by the addition of tests for critical conditions, for example, common shocks. To date, interpretation of the overall quality of a study has not benefitted from the use of a summative score for critical conditions.<sup>11</sup> Indeed, some critical conditions seem implicitly more relevant than others, such as the parallel trends and data requirements, critical conditions that many of the included studies in fact adhered to. The checklist could be accordingly improved, for example by implementing a hierarchy of relevance of critical conditions or exclusion criteria if a critical condition is violated. Additionally, uniform terminology within a DiD framework might help mitigate errors due to misinterpretation. As mentioned earlier, the Ryan *et al* checklist is not exhaustive, as a range of additional robustness and sensitivity checks exist. New methodological extensions of DiD are under development and specifications of DiD study designs can also be added, for example, synthetic control method, staggered treatment timing and heterogeneous treatment effects.<sup>14 46</sup>



Future research concerning the evaluations of early-life policy interventions, as well as other health policy interventions that apply DiD methods, could directly benefit from transdisciplinary collaboration. For example, numerous DiD designs have been further developed in econometrics,<sup>14</sup> but to date the uptake of these methodological extensions is limited in other research fields that also increasingly apply DiD. Transdisciplinary collaboration may help bridge the gap between the methodologically refined DiD techniques found in econometrics and the empirical perspective of policymakers and providers of health policy evaluations.

## Conclusions

DiD is increasingly applied to evaluate a broad range of early-life policy interventions, currently one of the most significant areas of improvement in health and social development. High-quality evaluations are therefore crucial for evidence-based policy-making concerning this critical period in life. This study revealed substantial variation in the current methodological quality of DiD studies on early-life policy interventions, with wide differences between studies as regards reporting and adherence to the proposed critical conditions of DiD. The fact that none of the included studies fully reported or adhered to the seven proposed criteria, is concerning as the validity of study conclusions depends on the extent to which critical conditions are met. To address this, we propose that a formal framework should be developed to facilitate unambiguous terminology and to assign a hierarchy of importance to critical conditions. Availability, acceptance and use of this type of DiD framework would undoubtedly contribute to improvements in the reporting and methodological quality of studies evaluating early-life policy interventions and result in improved policy decisions based on reliable evidence.

X Eline de Vries @eline\_f

**Acknowledgements** We thank the librarian Floor Boekelman for her assistance in setting up the search strategy. We also thank Albert Wong for his valuable input on the assessment of the critical conditions of difference-in-differences and Ardine de Wit and Marije Oosterhoff for their feedback on this work.

**Contributors** AK and JS conceptualised the study; AK and ML collected data under supervision from JS and EdV; AK, JS and EdV performed and reviewed the analysis; AK, JS and EdV wrote the initial draft of the manuscript; AP and MN reviewed and edited the manuscript. All authors read and approved the final version of the manuscript. AK is the guarantor.

**Funding** This study was part of the monitor bundled payment in maternity care that was funded by the Dutch Ministry of Health, Welfare and Sport and conducted by the National Institute for Public Health and the Environment (project numbers V/010038/01 and V/060438/22). The Ministry of Health, Welfare and Sport had no role in the design and execution of this study.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** All data relevant to the study are included in the article or uploaded as supplementary information.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

## ORCID iD

Anouk Klootwijk <http://orcid.org/0000-0003-4070-3424>

## REFERENCES

- Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* 2014;312:2401–2.
- Ashenfelter O. Estimating the effect of training programs on earnings. *Rev Econ Stat* 1978;60:47.
- Angrist JD, Pischke J-S. Mostly harmless Econometrics. In: *Mostly Harmless Econometrics*. Princeton University Press, Available: <https://www.degruyter.com/document/doi/10.1515/9781400829828/html>
- Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health* 2018;39:453–69.
- Saeed S, Moodie EEM, Strumpf EC, et al. Evaluating the impact of health policies: using a difference-in-differences approach. *Int J Public Health* 2019;64:637–42.
- Guan A, Hamad R, Batra A, et al. The revised Wic food package and child development: A quasi-experimental study. *Pediatrics* 2021;147:e20201853.
- Drewry J, Sen B, Wingate M, et al. The impact of the state children's health insurance program's unborn child ruling expansions on foreign-born Latina prenatal care and birth outcomes, 2000–2007. *Matern Child Health J* 2015;19:1464–71.
- Indrio F, Dargenio VN, Marchese F, et al. The importance of strengthening mother and child health services during the first 1000 days of life: the foundation of optimum health. *J Pediatr* 2022;245:254–6.
- United Nations. Transforming our world: The 2030 Agenda for Sustainable Development, Available: <https://sdgs.un.org/goals>
- Puinean G, Gokiart R, Taylor M, et al. Evaluation in the field of early childhood development: A Scoping review. *Evaluat J Australasia* 2022;22:63–89.
- Ryan AM, Burgess JF, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv Res* 2015;50:1211–35.
- Lechner M. The estimation of causal effects by difference-in-difference Methodsestimation of spatial panels. *FNT in Econometrics* 2010;4:165–224.
- Zeldow B, Hatfield LA. Confounding and regression adjustment in difference-in-differences studies. *Health Serv Res* 2021;56:932–41.
- Roth J, Sant'Anna PHC, Bilinski A, et al. What's trending in difference-in-differences? A synthesis of the recent Econometrics literature. *J Economet* 2023;235:2218–44.
- Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- World Bank. Country and Lending Groups World Bank, 2016. Available: [https://datahelpdesk.worldbank.org/knowledgebase/articles/906519#High\\_income](https://datahelpdesk.worldbank.org/knowledgebase/articles/906519#High_income)
- Brown CC, Moore JE, Felix HC, et al. Association of state Medicaid expansion status with low birth weight and Preterm birth. *JAMA* 2019;321:1598–609.
- Karimi M, Tsiachristas A, Looman W, et al. Bundled payments for chronic diseases increased health care expenditure in the Netherlands, especially for Multimorbid patients. *Health Policy* 2021;125:751–9.



- 19 Recio Alcaide A, Arranz JM. An impact evaluation of the strategy for normal birth care on Caesarean section rates and perinatal mortality in Spain. *Health Policy* 2022;126:24–34.
- 20 Lechner M. The estimation of causal effects by difference-in-difference methods. *FNT in Econometrics* 2010;4:165–224.
- 21 Freyaldenhoven S, Hansen C, Shapiro JM. Pre-event trends in the panel event-study design. *Am Eco Rev* 2019;109:3307–38.
- 22 Abbring JH, van den Berg GJ. The Nonparametric identification of treatment effects in duration models. *Econometrica* 2003;71:1491–517.
- 23 Sianesi B. An evaluation of the Swedish system of active labor market programs in the 1990s. *Rev Econom Stat* 2004;86:133–55.
- 24 Chay KY, Mcewan PJ, Urquiola M. The central role of noise in evaluating interventions that use test scores to rank schools. *Am Econ Rev* 2005;95:1237–58.
- 25 Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Associat* 1980;75:591.
- 26 Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press, Available: <https://www.cambridge.org/core/product/identifier/9781139025751/type/book>
- 27 Wherry LR, Fabi R, Schickedanz A, et al. State and Federal coverage for pregnant immigrants: prenatal care increased, no change detected for infant health. *Health Aff (Millwood)* 2017;36:607–15.
- 28 Hamad R, Collin DF, Baer RJ, et al. Association of revised WIC food package with perinatal and birth outcomes: A quasi-experimental study. *JAMA Pediatr* 2019;173:845–52.
- 29 Pulvera R, Collin DF, Hamad R. The effect of the 2009 WIC revision on maternal and child health: A quasi-experimental study. *Paediatr Perinat Epidemiol* 2022;36:851–60.
- 30 Montoya-Williams D, Passarella M, Lorch SA. The impact of paid family leave in the United States on birth outcomes and mortality in the first year of life. *Health Serv Res* 2020;55 Suppl 2:807–14.
- 31 Pihl AM, Basso G. Did California paid family leave impact infant health? *J Policy Anal Manage* 2019;38:155–80.
- 32 Häggström J, Sampaio F, Eurenus E, et al. Is the Salut programme an effective and cost-effective universal health promotion intervention for parents and their children? A register-based retrospective observational study. *BMJ Open* 2017;7:e016732.
- 33 Pulkki-Brännström AM, Lindkvist M, Eurenus E, et al. The equity impact of a universal child health promotion programme. *J Epidemiol Community Health* 2020;74:605–11.
- 34 Cygan-Rehm K, Karbownik K. The effects of Incentivizing early prenatal care on infant health. *J Health Econ* 2022;83:S0167-6296(22)00032-7.
- 35 de Jonge HC, Lagendijk J, Saha U, et al. Did an urban perinatal health programme in Rotterdam, the Netherlands, reduce adverse perinatal outcomes? register-based retrospective cohort study. *BMJ Open* 2019;9:e031357.
- 36 Meinhofer A, Witman A, Maclean JC, et al. Prenatal substance use policies and newborn health. *Health Economics* 2022;31:1452–67.
- 37 Cattan S, Farquaharson CG, Ginja R, et al. The health effects of universal early childhood interventions: evidence from sure start (working paper). 2021.
- 38 Rossin M. The effects of maternity leave on children's birth and infant health outcomes in the United States. *J Health Econ* 2011;30:221–39.
- 39 Janevic T, Hutcheon JA, Hess N, et al. Evaluation of a Multilevel intervention to reduce Preterm birth among black women in Newark, New Jersey: A controlled interrupted time series analysis. *Matern Child Health J* 2018;22:1511–8.
- 40 Dahlen HM, McCullough JM, Fertig AR, et al. Texas Medicaid payment reform: fewer early elective deliveries and increased gestational age and birthweight. *Health Aff (Millwood)* 2017;36:460–7.
- 41 Clarke D, Méndez GC, Sepúlveda DV. Growing together: assessing equity and efficiency in a Prenatal health program. *J Popul Econ* 2020;33:883–956.
- 42 Swartz JJ, Hainmueller J, Lawrence D, et al. Expanding prenatal care to unauthorized immigrant women and the effects on infant health. *Obstet Gynecol* 2017;130:938–45.
- 43 Davies NM, Smith GD, Windmeijer F, et al. Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology* 2013;24:363–9.
- 44 Hilton Boon M, Craig P, Thomson H, et al. Regression discontinuity designs in health: A systematic review. *Epidemiology* 2021;32:87–93.
- 45 von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *International Journal of Surgery* 2014;12:1495–9.
- 46 Kreif N, Grieve R, Hangartner D, et al. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ* 2016;25:1514–28.