



Passend bewijs voor preventie

Adviescommissie richtlijn passend bewijs
voor preventie

15 oktober 2025



Passend bewijs voor preventie

Rapport, 15 oktober 2025

Adviescommissie richtlijn passend bewijs voor preventie

Colofon

Leden van de Adviescommissie passend bewijs voor preventie

- Prof. dr. Bas ter Weel (voorzitter), SEO Economisch Onderzoek en Universiteit van Amsterdam
- Prof. dr. Jochen Mierau, Universitair Medisch Centrum Groningen
- Prof. dr. Ardine de Wit, Rijksinstituut voor Volksgezondheid en Milieu en Vrije Universiteit Amsterdam
- Dr. Marc Pomp, zelfstandig onderzoeker
- Dr. Bram Wouterse, Radboudumc
- Dr. Djoeke van Dale, Rijksinstituut voor Volksgezondheid en Milieu
- Drs. Heleen de Vaan, Zorginstituut Nederland

Adviseurs

- Dr. Roel van Elk, namens het Centraal Planbureau
- Prof. dr. Karien Stronks, namens de Gezondheidsraad

Waarnemers

- Drs. Rudo van den Brink, namens het ministerie van Volksgezondheid, Welzijn en Sport
- Drs. Bart van de Gevel, namens het ministerie van Financiën

Secretariaat

- Dr. Wouter Vermeulen (secretaris), SEO Economisch Onderzoek
- Kim van Berkel, MSc, SEO Economisch Onderzoek

Klankbordgroep

- Drs. Piet van den Reijen (Sociaal-Economische Raad)
- Jet Goppel, MSc, en Maarten Beijer, MSc (ZonMw)
- Drs. Nynke van Zorge (Vereniging van Nederlandse Gemeenten)
- Drs. Ton Coenen (GGD GHOR Nederland) en Sebastiaan Baan, MA (GGD West-Brabant)
- Marije de Leur, MSc (Zorgverzekeraars Nederland)

In opdracht van

Ministerie van Volksgezondheid, Welzijn en Sport

Voorwoord

De Adviescommissie richtlijn passend bewijs preventie (commissie) is op 1 april 2025 door de staatssecretaris van Volksgezondheid, Welzijn en Sport ingesteld (Kamerstukken II, 2024-25, 32793, nr. 819). De commissie heeft tot taak een richtlijn te ontwikkelen waarmee een passend niveau van bewijs kan worden vastgesteld waarmee het gezondheidseffect van een preventiemaatregel kan worden bepaald.

De gevraagde richtlijn is door de adviescommissie opgesteld in de periode april-oktober 2025. In deze periode heeft de commissie de voor het passend bewijs relevante bewijsmethoden in kaart gebracht. Dit is gebeurd aan de hand van literatuurstudie, ervaringen en richtlijnen op verschillende wetenschappelijke terreinen en in andere landen, interviews met experts en gesprekken met stakeholders. Ook is de commissie zeven keer bijeengekomen om te werken aan het advies.

Dit rapport is een technisch rapport. Het is bedoeld voor onderzoekers en beleidsmakers die interventies en beleidsmaatregelen beoordelen op kosten en baten, onderzoek uitvoeren naar effectief beleid of beleidsmaatregelen (willen) invoeren. Het rapport bouwt voort op de conclusies van het rapport *Preventie op waarde schatten* van de technische werkgroep Kosten en baten van preventie waarin het probleem van vaak ontoereikende methodologie wordt geconstateerd (TWG-KBP, 2023).

De eerste conclusie van de Adviescommissie richtlijn passend bewijs preventie is dat er ongeacht de maatregel en de context sprake moet zijn van overtuigend bewijs van de effectiviteit van een maatregel in de doelgroep waarop de maatregel zich richt. De adviescommissie heeft daarom een duidelijke ondergrens vastgesteld waar het gaat om het vaststellen van effectiviteit. Deze ondergrens kan als volgt worden samengevat: passend bewijs vereist minimaal inzicht in wat de uitkomst zou zijn zonder de maatregel. Dat betekent dat in elk geval duidelijk moet zijn welke hypothetische uitkomst zou zijn opgetreden als een bepaalde interventie of beleidsmaatregel niet had plaatsgevonden. Passend bewijs vraagt dus naast een interventiegroep om een controle(groep). Dit geldt voor alle maatregelen, ook als deze meervoudig of complex zijn. Deze ondergrens is bepaald aan de hand van de stand van de wetenschappelijke literatuur en geïllustreerd aan de hand van voorbeelden die laten zien dat deze ondergrens in de praktijk in uiteenlopende domeinen haalbaar is.

De tweede conclusie van de adviescommissie is dat veel studies in een andere context zijn uitgevoerd en niet direct vergelijkbaar zijn met de context waar de beleidsbeslissing betrekking op heeft. Dat stelt eisen aan het beoordelen van de toepasbaarheid, nadat vastgesteld is dat het bewijs van effectiviteit passend is. Dit vraagt in deze gevallen om een nauwgezette analyse van de context, van de werkzame elementen en mogelijkheden tot implementatie waar het gaat om interventies.

De derde conclusie van de adviescommissie is dat het vaststellen van langetermijneffecten onderdeel is van passend bewijs. Veel interventies hebben pas op langere termijn meetbare effecten waardoor een beoordeling van interventies op effectiviteit op korte termijn vaak negatief uitvalt. Een probleem hierbij is dat er dikwijls geen tijd is om langetermijneffecten af te wachten, zeker niet wanneer effecten zich over een periode van decennia manifesteren. De commissie adviseert daarom om beschikbare en betrouwbare intermediaire uitkomstmaten te gebruiken die indicatief zijn voor langetermijneffecten. Deze uitkomstmaten moeten beschreven zijn in de wetenschappelijke literatuur en onderdeel zijn van de effectmeting.

De richtlijn heeft verschillende gevolgen voor onderzoek en beleidsvorming op het terrein van preventie. Ten eerste leidt een minimale eis aan de bewijslast waarschijnlijk gemiddeld tot hogere

kosten voor het uitvoeren van onderzoek. De adviescommissie adviseert meer middelen voor onderzoek waarin van een “tegenfeitelijke situatie” wordt uitgegaan. Ten tweede vergt de richtlijn extra tijd van beleidsmakers, omdat de minimale eisen aan passend bewijs veranderen. Voor nieuwe maatregelen lijkt daarom een overgangperiode redelijk om te kunnen voldoen aan de redeneerlijn in dit advies. Ten slotte is experimenteren met nieuwe interventies en het monitoren van het gebruik van deze richtlijn van belang voor het vergaren van passend bewijs. De wetenschappelijke kennisbasis verbetert continu, wat monitoring en mogelijk bijstelling van de richtlijn vraagt.

Deze adviescommissie heeft een nieuwe richtlijn voor passend bewijs bij preventiemaatregelen ontwikkeld. Deze richtlijn helpt beleidsmakers, onderzoekers en uitvoeringsorganisaties om het juiste bewijsniveau voor de effectiviteit van preventieve interventies vast te stellen. Het RIVM bouwt hierop voort met een afwegingskader waarmee nieuwe maatregelen systematisch kunnen worden beoordeeld. In een volgende fase worden ook de budgettaire effecten van preventiebeleid inzichtelijk gemaakt, zodat besluitvorming beter kan worden onderbouwd.

Gedurende het proces heeft een secretariaat de adviescommissie ondersteund. Wouter Vermeulen en Kim van Berkel hebben uitstekend werk verricht en samen met de commissieleden grote delen van dit rapport geschreven. Ook hebben zij met een klankbordgroep van gedachten gewisseld en een aantal experts en stakeholders gesproken. We danken iedereen hartelijk voor de constructieve bijdrage aan het werk van deze commissie. Ten slotte was het een genoegen om in de commissie met bevlogen, ervaren en pragmatische wetenschappers samen te werken aan deze richtlijn.

Amsterdam, 15 oktober 2025

Prof. dr. Bas ter Weel

Voorzitter Adviescommissie richtlijn passend bewijs preventie

Samenvatting

De richtlijn voor passend bewijs voor preventie moet leiden tot een onderbouwing van kosten en baten van preventiemaatregelen die bruikbaar is voor besluitvorming over publieke uitgaven. Naast gerandomiseerd onderzoek kunnen alternatieve bewijsmethoden hiervoor toereikend zijn, mits gebruik wordt gemaakt van een controlegroep. Een verantwoorde toepassing van bewijs vereist dat de onderzoekscontext en de uitvoeringspraktijk vergelijkbaar zijn en de werkzaamheid van de maatregel theoretisch onderbouwd is. Passend bewijs voor langetermijneffecten kan gebaseerd zijn op intermediaire effecten en modellering. Voor gebruik van de richtlijn geldt het ‘pas toe of leg uit’-principe.

Dit rapport bevat een richtlijn waarmee een passend bewijsniveau kan worden vastgesteld om het gezondheidseffect van een preventiemaatregel te bepalen. Binnen de gezondheidszorg wordt voor het vaststellen van bewijs vaak sterk uitgegaan van gerandomiseerde onderzoeken (randomized controlled trials (RCT's)) en systematische reviews daarvan, maar het is voor lang niet alle preventiemaatregelen redelijkerwijs haalbaar en ook niet noodzakelijk om RCT's uit te voeren. Deze richtlijn geeft beleidsmakers, onderzoekers en kennisinstituten daarom houvast voor alternatieve bewijsmethoden die passend bewijs kunnen opleveren.

Bewijsmethoden

Passend bewijs vereist minimaal een overtuigende benadering van de “tegenfeitelijke situatie”, oftewel de hypothetische uitkomst die zou zijn opgetreden als een bepaalde interventie of beleidsmaatregel niet had plaatsgevonden. De effectbepaling bestaat uit een vergelijking van de waargenomen uitkomst met deze tegenfeitelijke uitkomst. Randomisatie is hiervoor niet altijd noodzakelijk, maar passend bewijs vraagt wel altijd om onderzoek dat naast de interventiegroep ook een controlegroep betreft.

In de ordening van de Maryland Scientific Methods Scale (SMS) vereist passend bewijs een goede toepassing van bewijsmethoden vanaf niveau 3. Een veel toegepaste methode op dit niveau vergelijkt veranderingen in uitkomsten voor en na een interventie tussen groepen die de interventie wel en niet hebben ondergaan, zonder dat sprake is van randomisatie. Andere bewijsmethoden op ten minste dit niveau benutten bijvoorbeeld de stapsgewijze introductie van een interventie of een drempelwaarde in de toekenning van een interventie om tot een controlegroep te komen. De richtlijn bespreekt de voorwaarden waaronder deze en andere bewijsmethoden tot een overtuigende benadering van de tegenfeitelijke situatie leiden om effectiviteit vast te stellen.

De inzet van bewijsmethoden op ten minste SMS-niveau 3 is op uiteenlopende preventiedomeinen haalbaar. Dit blijkt uit verschillende voorbeelden van onderzoek naar preventieaanpakken op het gebied van leefstijl, leefomgeving, werk, mentale gezondheid en jeugd. Ook de effectiviteit van bijvoorbeeld accijnzen, regulering, vaccinaties en bevolkingsonderzoeken is aangetoond op ten minste dit niveau.

Toepasbaarheid

De toepassing van onderzoeksresultaten uit een andere context vereist dat de onderzoeks- en implementatiecontext vergelijkbaar zijn op kenmerken die van invloed zijn op de werkzaamheid van een preventiemaatregel. Dit is van belang omdat de effectiviteit van een maatregel in een ex-ante beoordeling doorgaans wordt bepaald op basis van onderzoek uit een andere context, zoals een andere gemeente, een ander land of een experimentele context. De richtlijn geeft een overzicht van kenmerken die hierbij een rol kunnen spelen.

Een goede inschatting van de toepasbaarheid van onderzoek vereist daarom ook zicht op de werkzame elementen van een aanpak. Een minimale vereiste hiervoor is een theoretische onderbouwing van waarom en hoe deze aanpak tot de beoogde resultaten moet leiden. Het rapport bespreekt verschillende kwalitatieve en kwantitatieve methoden om werkzame elementen empirisch vast te stellen.

Langetermijneffecten

Dat effecten van een maatregel pas op lange termijn optreden hoeft geen belemmering te zijn voor passend bewijs. Passend bewijs voor langetermijneffecten kan bijvoorbeeld gebaseerd zijn op intermediaire effecten en modellering. De eisen van de richtlijn met betrekking tot bewijsmethoden zijn dan van toepassing op het vaststellen van de intermediaire effecten van een preventiemaatregel. Voor de relatie tussen intermediaire uitkomsten en einduitkomsten moet ook een overtuigende theoretische en empirische onderbouwing bestaan. Als langetermijneffecten met een model worden bepaald, dan moet dit model empirisch gevalideerd zijn.

Onderzoek dat de langetermijneffecten van een preventiemaatregel bepaalt, moet transparant zijn over de gemaakte aannames en de onzekerheden die hieruit volgen. Het gaat bijvoorbeeld om onzekerheden over deelname, duur van het effect en gedragsreacties van deelnemers en andere actoren. Dit kan aan de hand van gevoeligheidsanalyses. Modelstudies moeten altijd gevoeligheidsanalyses bevatten.

Gebruik van de richtlijn en aanbevelingen

Voor gebruik van de richtlijn geldt het 'pas toe of leg uit'-principe. Daarbij mag de richtlijn experimenten niet in de weg staan die niet direct passend bewijs opleveren, maar wel tot stappen in deze richting leiden. Zo moet er ruimte blijven voor de ontwikkeling van nieuwe aanpakken. Verder is het belangrijk om het gebruik van de richtlijn te monitoren en na een aantal jaren te evalueren.

Onderzoeksbekostiging en data-infrastructuur zijn randvoorwaardelijk voor een succesvolle toepassing van de richtlijn. Bij de verdeling van onderzoeksgelden moet rekening worden gehouden met de kosten van het betrekken van een controlegroep, zodat gecorrigeerd wordt voor het feit dat dit bij sommige aanpakken moeilijker te realiseren is. Een verbeterde data-infrastructuur vergroot de mogelijkheden voor onderzoek op basis van niet-experimentele methoden. Dit vereist nader onderzoek naar wat hiervoor nodig is op het gebied van de verzameling en ontsluiting van gegevens.

Summary

The guideline on the sufficient level of evidence for prevention should lead to a substantiation of the costs and benefits of preventive measures, which can be used for decision-making on public expenditure. In addition to randomized research, alternative research methods may also be sufficient for this purpose, provided that a control group is used. The responsible application of evidence requires that the research context and implementation practices are comparable, and that the effectiveness of the measure is theoretically substantiated. Sufficient evidence for long-term effects can be based on intermediate effects and modeling. The “comply or explain” principle applies to the use of the guideline.

This report contains a guideline for determining the sufficient level of evidence to assess the health impact of preventive measures. In healthcare, evidence is often primarily based on randomized controlled trials (RCTs) and their systematic reviews. However, for many preventive measures, conducting RCTs is neither feasible nor necessary. This guideline offers policymakers, researchers, and knowledge institutes guidance on alternative research methods that can yield sufficient evidence.

Methods of Evidence

Sufficient evidence requires at least a convincing approximation of the “counterfactual situation,” which is the hypothetical outcome that would have occurred if a certain intervention or policy measure had not taken place. The effect is determined by comparing the observed outcome with this counterfactual outcome. This does not always require randomization, but sufficient evidence does require research that includes both an intervention group and a control group.

Sufficient evidence requires sound application of research methods on at least level 3 Maryland Scientific Methods Scale (SMS). A frequently used method at this level that does not use randomization compares changes in outcomes before and after an intervention between groups that did and did not receive the intervention. Other methods on at least this level include, for example, the stepwise introduction of an intervention or the use of a threshold in the allocation of an intervention to establish a control group. The guideline discusses the conditions under which these and other methods can lead to a convincing approximation of the counterfactual situation in order to establish effectiveness.

The use of research methods on at least SMS level 3 is feasible in various prevention domains. This is demonstrated by several examples of studies on prevention approaches in the areas of lifestyle, living environment, work, mental health, and youth. The effectiveness of measures such as excise duties, regulation, vaccination, and screening programs has also been demonstrated at least at this level.

Applicability

Applying research results from another context requires that the research and implementation contexts are comparable in characteristics that influence the effectiveness of a preventive measure. This is important because the effectiveness of a measure in an ex-ante assessment is usually based on research conducted in another context, such as a different municipality, country, or experimental setting. The guideline provides an overview of the relevant characteristics involved.

A good assessment of applicability also requires insight into the active components of an approach. A minimum requirement is a theoretical substantiation of why and how the approach should lead to

the intended outcomes. The report discusses various qualitative and quantitative methods to empirically determine the active components.

Long-term Effects

The fact that effects of a measure may only manifest in the long term does not have to be an obstacle to obtaining sufficient evidence. Sufficient evidence for long-term effects can, for example, be based on intermediate effects and modelling. The guideline's requirements regarding research methods apply to establishing the intermediate effects of a preventive measure. For the relationship between intermediate and final outcomes, both a convincing theoretical and empirical substantiation must exist. When long-term effects are determined using a model, that model must be empirically validated.

Research assessing the long-term effects of a preventive measure must be transparent about its assumptions and the resulting uncertainties. These may include uncertainties about participation, effect duration, and behavioural responses from participants and other actors. Such uncertainties can be addressed through sensitivity analyses. Modelling studies must always include sensitivity analyses.

Use of the Guideline and Recommendations

The "comply or explain" principle applies to the use of this guideline. The guideline should not hinder experiments that may not immediately produce appropriate evidence but do contribute to progress in that direction. There should always be room for developing new approaches. It is also important to monitor the use of the guideline and evaluate it after several years.

Adequate research funding and data infrastructure are essential preconditions for the successful application of the guideline. The allocation of research funds should take into account the costs associated with including a control group and adjust for the fact that this may be more difficult in some approaches. An improved data infrastructure enhances the possibilities for research based on non-experimental methods. This requires further study of what is needed in terms of data collection and accessibility.

Inhoudsopgave

Colofon	2
Voorwoord	3
Samenvatting	5
Summary	7
Inhoudsopgave.....	9
1 Inleiding.....	10
1.1 Achtergrond	10
1.2 Doel en bereik van deze richtlijn.....	11
1.3 Opbouw van de richtlijn.....	12
1.4 Werkwijze van de commissie	13
1.5 Dit rapport.....	13
2 Aansluiting bij bestaande richtlijnen.....	14
2.1 Passend bewijs in een MKBA	14
2.2 Passend bewijs voor verzekerde zorg	15
2.3 Passend bewijs volgens de Gezondheidsraad.....	17
2.4 Erkende interventies	18
2.5 Conclusie.....	18
3 Bewijsmethode.....	19
3.1 SMS-niveau 5	21
3.2 SMS-niveau 4	23
3.3 SMS-niveau 3	26
3.4 SMS-niveaus 1 en 2	28
3.5 Voorbeelden	29
4 Toepasbaarheid.....	32
4.1 Context en werkzame elementen.....	33
4.2 Wat vraagt dit van het onderzoek?.....	35
5 Langetermijneffecten	37
5.1 Bronnen van onzekerheid	38
5.2 Intermediaire uitkomsten	39
5.3 Empirische validatie	40
5.4 Onzekerheid	41
6 Aanbevelingen.....	43
Bronnenlijst	46
Bijlage A Gesprekspartners	53
Bijlage B Voorbeelden.....	54

1 Inleiding

Bij het vaststellen van bewijs ligt de nadruk in de gezondheidszorg op experimenteel onderzoek, maar voor preventie is dat niet altijd mogelijk. Welke andere bewijsmethoden leveren ook passend bewijs? Deze richtlijn geeft onderzoekers, beleidsmakers en kennisinstituten houvast.

1.1 Achtergrond

Beperkte inzet publieke middelen op preventie

Door investeringen in preventie leven mensen langer en gezonder. Sinds de jaren zeventig leven we in Nederland bijna tien jaar langer en de levensverwachting in als goed ervaren gezondheid nam sinds de jaren tachtig met vijf jaar toe (CBS, 2025a). Preventie speelt hierin een belangrijke rol. Zo heeft anti-rookbeleid eraan bijgedragen dat er tegenwoordig aanzienlijk minder mensen sterven aan hart- en vaatziekten.¹ Een overzichtsstudie van preventief gezondheidsbeleid door het RIVM vindt ook sterke aanwijzingen voor gezondheidseffecten van alcoholpreventie (RIVM, 2021). Daarnaast hebben maatregelen op bijvoorbeeld het gebied van luchtkwaliteit, verkeersveiligheid, arbowetgeving en vaccinatie bijgedragen aan een hogere levensverwachting.

Sommige preventiemaatregelen leveren per euro veel meer op aan gezonde levensverwachting dan uitgaven aan curatieve zorg. De Wetenschappelijke raad voor het regeringsbeleid (WRR) somt een aantal voorbeelden op van kosteneffectieve preventiemaatregelen, waaronder het verbod op tv-reclame voor ongezond eten gericht op kinderen (WRR, 2021).² Deze maatregel kost drie euro per gewonnen levensjaar in goede gezondheid (QALY). In de curatieve zorg worden hiervoor referentiewaarden van 20 tot 80 duizend euro gebruikt (Zorginstituut Nederland, 2024a).³ Ook maatregelen zoals de screening door huisartsen op overmatige alcoholconsumptie of het verlagen van de maximumsnelheid in dichtbevolkte gebieden naar 80 kilometer per uur kosten aanzienlijk minder per QALY.

De overheid geeft echter aanzienlijk minder uit aan preventie dan aan curatieve zorg. Het ministerie van Volksgezondheid, Welzijn en Sport (VWS) geeft in 2024 naar schatting 1,8 miljard euro uit aan preventie.⁴ Het gaat om gezondheidsbeleid en gezondheidsbescherming, ziektepreventie, gezondheidsbevordering en sport en bewegen (Kamerstukken II, 2024-25, 32793 nr. 783). Daarentegen is in 2024 113,5 miljard euro uitgegeven aan zorg. Hiervan gaat meer dan de helft naar zorg die verzekerd is onder de Zorgverzekeringswet (Zvw) en doorgaans een curatief karakter heeft (CBS, 2025b).

Publieke investeringen in preventie vereisen passend bewijs

Om publieke middelen in preventie te investeren moet de effectiviteit van een preventiemaatregel voldoende zijn aangetoond. De Technische werkgroep kosten en baten van preventie (TWG-KBP, 2023) constateert dat binnen de gezondheidszorg voor het vaststellen van bewijs vaak sterk uit wordt gegaan van gerandomiseerde onderzoeken (randomized controlled trials (RCT's)) en systematische reviews daarvan, maar dat het voor lang niet alle preventiemaatregelen redelijkerwijs

¹ <https://www.cbs.nl/nl-nl/fag/specifiek/waardoor-stijgt-de-levensverwachting->

² Zie Van der Vliet et al. (2020) voor een breder overzicht van kosteneffectieve maatregelen in verschillende domeinen om de publieke gezondheid te verbeteren.

³ TWG-KBP (2023) beveelt een referentiewaarde van 50 duizend euro per QALY aan voor preventiebeleid.

⁴ Uitgaven aan preventie door andere departementen, zoals investeringen in verkeersveiligheid, blijven in deze vergelijking buiten beschouwing.

haalbaar is om RCT's uit te voeren. Dit komt bijvoorbeeld omdat randomisatie lastig of niet mogelijk is, er sprake is van langetermijneffecten die buiten de doorlooptijd van een RCT vallen of dat invoering van de preventiemaatregel alleen op landelijke schaal mogelijk is. Daarom adviseert de werkgroep om een raamwerk te ontwikkelen dat beleidsmakers, onderzoekers en kennisinstituten houvast geeft over passend bewijs.

Een richtlijn voor passend bewijs vormt ook de eerste pijler van een investeringsmodel voor preventie van het ministerie van VWS (Kamerstukken II, 2024-25, 32793, nr. 819). Andere pijlers in dit model gaan over het afwegingskader, de financiële besluitvorming en monitoring en evaluatie. Het investeringsmodel moet de beoordeling van preventie op gezondheidseffecten faciliteren en meer inzicht bieden in budgettaire gevolgen. Dit helpt om te komen tot goed onderbouwde besluitvorming binnen de gestelde begrotingsregels. Het is daarbij mogelijk om maatregelen (deels) te dekken met besparingen. De opname van het RS-vaccin in het rijksvaccinatieprogramma, waarbij kosten deels zijn gedekt met de besparing op Zvw-uitgaven die vaccinatie tegen het RS-virus oplevert, vormt hiervan een voorbeeld. Ook hebben verschillende zorgpartijen en overheden in het recent afgesloten Aanvullend Zorg en Welzijnsakkoord (AZWA) met elkaar afgesproken dat besparingen in de Zvw kunnen worden ingezet voor medische preventie en in het sociaal domein (Rijksoverheid, 2025). Duidelijkheid over passend bewijs is hiervoor een expliciete randvoorwaarde.

1.2 Doel en bereik van deze richtlijn

In navolging van het advies van de TWG-KBP heeft het kabinet de Adviescommissie richtlijn passend bewijs preventie ingesteld om de staatssecretaris van VWS te adviseren over het hanteren van een door de commissie te ontwikkelen richtlijn waarmee het passend niveau van bewijs kan worden vastgesteld voor de effectbepaling van preventiemaatregelen (VWS, 2025). Deze richtlijn is bedoeld voor beleidsambtenaren, kennisinstituten en onderzoeksbureaus die het bewijs voor de effectiviteit van preventie moeten beoordelen.

De richtlijn is van toepassing op al het preventiebeleid, zowel binnen als buiten de gezondheidszorg, dat zich primair of secundair richt op het behouden van een goede gezondheid of het voorkómen van het ontstaan of verergeren van gezondheidsproblemen. Hieronder valt dus niet alleen beleid waarin gezondheid centraal staat, zoals bij vaccinaties, bevolkingsonderzoeken of leefstijlinterventies, maar ook armoedebelief of maatregelen in de leefomgeving waarbij gezondheid een secundair doel is. Deze afbakening van preventie sluit aan bij het advies van de TWG-KBP.⁵

Beleidsmaatregelen kunnen allerlei vormen aannemen. Het kan bijvoorbeeld gaan om de inzet van een interventie door een gemeente, het opnemen van preventieve zorg in het basispakket, regulering van vervuilende uitstoot of een belasting op suiker in frisdrank. Interventies kunnen enkelvoudig zijn, maar het kan ook gaan om programma's zoals Kansrijke Start en de ketenaanpak voor jongeren met overgewicht en obesitas, waarbij interventies worden aangepast aan de lokale context. In dit rapport gebruiken we in navolging van het advies van de TWG-KBP doorgaans de term interventie, maar deze andere typen van preventiemaatregelen vallen hier dus ook onder.

De richtlijn moet ertoe leiden dat kosten en baten van preventiemaatregelen zodanig worden onderbouwd dat ze bruikbaar zijn voor besluitvorming over publieke uitgaven. Effectbepalingen zijn

⁵ Dit rapport geeft ook een nader overzicht van verschillende soorten preventie naar doelstelling (gezondheidsbescherming, gezondheidsbevordering en ziektepreventie), stadium van de aandoening waarop preventie zich richt (primair, secundair en tertiair) en doelgroep (universeel, selectief, geïndiceerd en zorggerelateerd). Al deze soorten van preventie vallen dus binnen het bereik van deze richtlijn.

voor de berekening van kosten en baten namelijk van doorslaggevend belang. Als een effect niet overtuigend is aangetoond, kan er ook geen betrouwbare uitspraak worden gedaan over de kosten of baten die hieruit volgen. De leidraad voor maatschappelijke kosten-batenanalyse (MKBA) stelt dan ook strenge eisen aan de empirische onderbouwing van de effecten van maatregelen (Romijn & Renes, 2013). De tweede pijler van het investeringsmodel van het ministerie van VWS zal nader duiden hoe kosten en baten in dit model berekend en afgewogen moeten worden.

Voor een kosten-batenafweging van preventiemaatregelen moet het passende bewijs betrekking hebben op de uitkomsten die aan de basis staan van de belangrijkste maatschappelijke baten. In het geval van preventiemaatregelen die primair of secundair gericht zijn op gezondheid gaat het dan dus om gezondheidsuitkomsten. Uitkomsten waarop de richtlijn betrekking heeft zijn echter niet beperkt tot gezondheid. Het kan ook gaan om bredere uitkomsten zoals arbeidsdeelname, zorgkosten, mantelzorg of welzijn. Het effect van preventiemaatregelen op deze uitkomsten kan namelijk ook van belang zijn om een volledig beeld van de maatschappelijke kosten en baten te geven. Ook deze effecten moeten dan passend onderbouwd zijn.

De richtlijn is geen dwangbuis. De aanbevelingen kunnen voor veel effectbepalingen worden gebruikt, maar er kunnen ook situaties optreden waarin een afwijkende aanpak beter is. Een keuze om af te wijken van de richtlijn vereist een goede inhoudelijke onderbouwing. De bewijslast ligt bij de (onderzoeks)partij die het bewijs voor de effectiviteit van een maatregel beoordeelt en die overtuigend moet onderbouwen waarom een afwijkende aanpak beter zou zijn. Uiteindelijk is het aan de politiek om te bepalen of deze afwijkende aanpak voldoende overtuigend is om een investering van publieke middelen te rechtvaardigen.

1.3 Opbouw van de richtlijn

Passend bewijs omvat zowel de interne validiteit van onderzoek als de toepassing hiervan op de preventiemaatregel waarover de overheid moet besluiten. Interne validiteit van het bewijs gaat over de zekerheid waarmee het onderzoek de oorzaak-gevolgrelatie tussen een maatregel en de beoogde uitkomsten binnen de context van het onderzoek vaststelt. De toepasbaarheid van intern valide bewijs is belangrijk omdat de overheid een maatregel doorgaans beoordeelt voordat die wordt ingevoerd. Hiervoor is dan bewijs nodig uit bijvoorbeeld een ander land, een andere gemeente of een experimentele context.

De richtlijn geeft een overzicht van bewijsmethoden die, mits goed toegepast, voldoende interne validiteit opleveren. Dit overzicht is gebaseerd op een ordening van bewijsmethoden volgens de Maryland Scientific Methods Scale (Farrington et al., 2002). Dit is een breed geaccepteerde ordening die zowel binnen als buiten het zorgdomein kan worden toegepast. De Maryland Scale gaat niet alleen in op de interne validiteit van RCT's ten opzichte van methoden op basis van observationele data, maar maakt ook binnen de observationele methoden onderscheid naar bewijskracht. Dit onderscheid is nodig om voor bewijsmethoden buiten de RCT's, die juist bij preventiemaatregelen niet altijd mogelijk of wenselijk zijn, aan te kunnen geven wanneer ze passend zijn.

Toepasbaarheid moet worden beoordeeld op de mate waarin verschillen tussen de onderzoeks- en implementatiecontext de werkzaamheid van een preventiemaatregel beïnvloeden. Dit illustreren we aan de hand van voorbeelden en we gaan in op wat dit vraagt van evaluatieonderzoek. Daarbij bouwen we voort op richtlijnen voor de evaluatie van complexe interventies (Pfadenhauer et al., 2017; Skivington et al., 2021). Kenmerkend voor dit soort interventies is dat ze bestaan uit meerdere werkzame elementen en dat de inzet hiervan wordt aangepast aan de lokale context of individuele

situatie. Juist binnen het preventiedomein wordt veel van dit soort interventies verwacht, terwijl de toepasbaarheid van onderzoek uit een andere context ingewikkeld is.

Daarnaast gaat de richtlijn in op langetermijneffecten. Als een preventiemaatregel op lange termijn (gezondheids)effecten heeft, dan moeten die meegenomen worden in een volledig beeld van de maatschappelijke kosten en baten. In een MKBA is dit verplicht. Daarom is effectbepaling op de lange termijn ook onderdeel van deze richtlijn. De richtlijn gaat in op het gebruik van intermediaire effecten en modellering om langetermijneffecten te bepalen en op de eisen die hieraan gesteld mogen worden. Daarbij ligt de nadruk op de empirische betrouwbaarheid en de manier waarop onzekerheden inzichtelijk gemaakt kunnen worden.

De richtlijn bespreekt twintig voorbeelden van onderzoek naar de effectiviteit van preventiemaatregelen op uiteenlopende terreinen. Hieruit blijkt dat er op al deze terreinen onderzoek bestaat dat voldoet aan de criteria van de richtlijn met betrekking tot interne validiteit. Deze criteria lijken daarmee dus niet alleen wenselijk maar ook haalbaar. De bespreking van deze voorbeelden gaan ook in op de toepasbaarheid van de resultaten buiten de onderzoekscontext.

1.4 Werkwijze van de commissie

De adviescommissie heeft de richtlijn opgesteld op basis van literatuurstudie, bestaande richtlijnen op verschillende wetenschappelijke terreinen in Nederland en in andere landen en op de eigen ervaring van commissieleden. Daarnaast zijn er interviews gehouden met diverse experts en stakeholders en heeft er een expertbijeenkomst plaatsgevonden, zie Bijlage A voor een overzicht. Inzichten hieruit zijn ook meegenomen bij het opstellen van de richtlijn.

1.5 Dit rapport

Dit rapport bespreekt bewijsmethoden in Hoofdstuk 3, toepasbaarheid in Hoofdstuk 4 en langetermijneffecten in Hoofdstuk 5. Hoofdstuk 2 geeft een overzicht van bestaande richtlijnen voor passend bewijs voor preventie. Hier sluit de richtlijn van de adviescommissie zoveel mogelijk bij aan. Conclusies en aanbevelingen staan in Hoofdstuk 6. De uitgewerkte voorbeelden zijn opgenomen in Bijlage B.

2 Aansluiting bij bestaande richtlijnen

Er bestaan al verschillende richtlijnen die betrekking hebben op passend bewijs voor besluitvorming over publieke uitgaven aan preventie. Aansluiting bij deze richtlijnen is wenselijk, zodat er een consistent en breedgedragen beeld ontstaat. Dit hoofdstuk geeft een overzicht.

De richtlijn van de adviescommissie sluit zoveel mogelijk aan bij bestaande richtlijnen die betrekking hebben op passend bewijs voor besluitvorming over publieke uitgaven aan preventie. In dit hoofdstuk gaan we in op de werkwijzer voor maatschappelijke kosten-batenanalyse (MKBA) in het sociaal domein, de eisen die de Zorgverzekeringswet stelt met betrekking tot passend bewijs en de werkwijze van de Gezondheidsraad voor de beoordeling van vaccinaties en bevolkingsonderzoeken. Daarnaast bespreken we de criteria voor opname in een van de databases van erkende interventies.

Het gebruik van de werkwijzer voor MKBA's is verplicht voor alle MKBA's in opdracht van de Rijksoverheid.⁶ Een MKBA is echter geen advies, maar een hulpmiddel voor beleidsmakers bij het maken van beslissingen over publieke investeringen. Zorginstituut Nederland (ZIN) en de Gezondheidsraad adviseren de overheid wel. Bij ZIN gaat het over opnemen van zorg in het pakket van verzekerde zorg. De Gezondheidsraad adviseert onder meer over het opnemen van vaccinaties en bevolkingsonderzoeken in door de Rijksoverheid gefinancierde programma's. Erkenning van een interventie staat los van beslissingen over publieke uitgaven en heeft enkel ten doel om inzicht te geven in de werkzaamheid ervan.

2.1 Passend bewijs in een MKBA

Een MKBA brengt kosten en opbrengsten systematisch in kaart en drukt deze posten waar mogelijk uit in geld. Deze benadering wordt toegepast op een breed scala van beleidsmaatregelen. Er is een algemene leidraad die vastlegt waaraan een MKBA moet voldoen (Romijn en Renes, 2013) en er bestaan allerlei verbijzonderingen van deze leidraad voor toepassingsgebieden. Er is ook een werkwijzer voor MKBA in het sociaal domein (Koopmans et al., 2016) waar het leeuwendeel van de maatregelen op het gebied van preventie onder valt.

Het bewijsmateriaal voor de belangrijkste effecten in een MKBA in het sociaal domein moet volgens deze werkwijzer gebaseerd zijn op sociale experimenten op basis van loting, vergelijkbaar met een RCT, of op observationele methoden die gebruikmaken van quasi-experimentele variatie. Als observationele methoden noemt de werkwijzer *regression discontinuity*, *difference-in-differences* en instrumentele variabelen; het volgende hoofdstuk licht deze methoden nader toe. Als effecten worden bepaald op basis van wetenschappelijk onderzoek dat deze methoden gebruikt in een andere context, dan moet dit onderzoek toepasbaar zijn op de onderzochte interventie. Verder voldoen gevalideerde modellen aan deze standaard.

Bij de indeling van bewijsmethoden baseert de werkwijzer zich op een rapport van de Expertwerkgroep Effectmeting (Commissie Theeuwes, 2012). Deze expertwerkgroep heeft gestructureerd evaluatiemogelijkheden in beeld gebracht om de directe effecten van een aantal beleidsmaatregelen van het ministerie van Economische Zaken te meten. De expertwerkgroep

⁶ Er is ook een algemene richtlijn voor de beoordeling van nieuwe beleidsvoorstellen van de Rijksoverheid, namelijk het [Beleidskompas](#). Een van de vragen die hierin beantwoord moet worden gaat over de effecten van beleid. De MKBA is een van de analysemethoden die deze richtlijn aanbeveelt. Koopmans et al. (2024) gaan in op de relatie tussen MKBA en het Beleidskompas.

brengt hierbij ook een ordening aan in de overtuigingskracht van verschillende experimentele en observationele methoden. Deze ordening sluit nauw aan bij de Maryland Scale.

Er bestaan ook aan de MKBA gerelateerde beoordelingsinstrumenten, zoals een Kengetallen KBA of een Indicatieve KBA, waarvoor volgens de werkwijzer voor MKBA in het sociaal domein een lichtere bewijslast geldt. Het gaat dan bijvoorbeeld om *propensity score matching* of om keuze-experimenten. Met deze lichtere bewijsmethoden kunnen maatschappelijke kosten en baten nog steeds indicatief in kaart worden gebracht, maar het onderscheid tussen MKBA's en deze lichtere beoordelingsinstrumenten maakt het voor beleidsmakers duidelijk dat de onderbouwing dan minder stevig is.

De werkwijzer voor MKBA's in het sociaal domein raadt ook enkele methoden van bewijsvoering af, namelijk effectschattingen baseren op verwachtingen van betrokken actoren en veronderstellen dat beleidsdoelen zullen worden behaald. Deze aanbeveling moeten we in het licht zien van een andere constatering van de werkwijzer, namelijk dat een gebrekkige kwaliteit van bewijsmateriaal juist in het sociaal domein een punt van zorg is. In een inventarisatie van onderzoek rond arbeidsmarkt en sociale zekerheid constateren Koning et al. (2013) dat veel studies een minder geloofwaardig onderzoeksdesign hebben waardoor de effecten sterk vertekend kunnen zijn. Ook zijn gebruikte wetenschappelijke studies vaak niet goed toepasbaar op de maatregel die in de MKBA wordt onderzocht.

2.2 Passend bewijs voor verzekerde zorg

Voor opname in het basispakket van verzekerde zorg gelden vier pakketcriteria: effectiviteit, noodzakelijkheid, kosteneffectiviteit en uitvoerbaarheid (Zorginstituut Nederland, 2023a). Het pakketcriterium 'effectiviteit' houdt in dat de zorg bewezen effectief moet zijn, en komt overeen met het in de Zorgverzekeringswet (Zvw) en de Wet langdurige zorg (Wlz) opgenomen wettelijke criterium 'stand van de wetenschap en praktijk'.

Een groot deel van de verzekerde zorg is in de Zvw generiek omschreven waardoor deze veelal zonder expliciete toetsing vooraf in het basispakket kan instromen (het zogeheten 'generiek omschreven' of 'open' systeem). Dit geldt onder andere voor medisch-specialistische zorg. Deze moet voldoen aan de wettelijke criteria 'zorg zoals zorgprofessionals plegen te bieden' en aan de 'stand van de wetenschap en praktijk' (SWP) (art. 2.4 lid 1 Besluit zorgverzekering; art. 2.1 lid 2 Besluit zorgverzekering).⁷ Zorgverzekeraars en zorgprofessionals kunnen zorg aan deze criteria toetsen. Pas als zij er niet uitkomen kan het Zorginstituut (ZIN) gevraagd worden een 'duiding' te doen. ZIN toetst de zorg dan aan het 'plegen te bieden' criterium en aan de SWP, en geeft in een *standpunt* aan of de zorg voldoet aan deze Zvw-criteria.

Over nieuwe zorg in het 'gesloten systeem' (zoals extramurale geneesmiddelen en paramedische zorg) beoordeelt het Zorginstituut de effectiviteit (SWP) én er wordt getoetst aan de andere drie pakketcriteria. De uitkomst is een *advies* van het Zorginstituut aan de minister van VWS. Indien deze besluit het advies te volgen wordt vervolgens de wet aangepast.

Preventieve zorg in het basispakket betreft altijd individuele (geïndiceerde) preventie. Aan de SWP getoetste preventieve interventies, die inmiddels opgenomen zijn in het basispakket, zijn bijvoorbeeld valpreventie, de gecombineerde leefstijlinterventie en het stoppen-met-rokenprogramma.

⁷ Door deze generieke omschrijving is voor instroom van deze zorg in het basispakket geen wetswijziging (en veelal ook geen toetsing aan de andere pakketcriteria) nodig.

Voor het beoordelen van de SWP heeft het Zorginstituut een beoordelingskader opgesteld (Zorginstituut Nederland, 2023b).⁸ Een SWP-beoordeling start met het formuleren van de doelstelling van de interventie (claim) en het opstellen van een beschrijving van de Populatie, Interventie, Comparator, Outcome, Tijd en Setting (PICO(ts)). Ook wordt beoordeeld welke onderzoeksmethodiek passend is voor het beantwoorden van de pakketvraag. Met passend onderzoek bedoelt ZIN dat de bewijskracht in overeenstemming moet zijn met de bewijslast. Met bewijskracht wordt de kwaliteit van het bewijs bedoeld, en met bewijslast wat nodig én haalbaar is gegeven de te beoordelen interventie-indicatiecombinatie. Een leidraad voor het inschatten van de noodzakelijke én haalbare onderzoekskarakteristieken is de zogenoemde passend onderzoeksvragenlijst (zie Heymans et al., 2013).

Bij de beoordeling van de SWP wordt gebruikgemaakt van gepubliceerde wetenschappelijke studies en worden de principes van *evidence-based medicine* gehanteerd. Er wordt geen gebruik gemaakt van een ordening van observationele bewijsmethoden zoals de Maryland Scale. De kwaliteit van bewijs wordt beoordeeld aan de hand van de GRADE-methode (*Grading of Recommendations, Assessment, Development and Evaluations*). Binnen deze methode is belangrijk dat voorafgaand aan de beoordeling overeenstemming bestaat over de belangrijkste uitkomsten die meegenomen worden in de beoordeling. De kern van GRADE is dat op systematische wijze bewijs gezocht en geselecteerd wordt over het effect dat de interventie heeft op die uitkomstmaten. Per uitkomst wordt het bewijs op transparante en gestructureerde wijze beoordeeld.

Een goed opgezette en uitgevoerde RCT geeft volgens deze systematiek a priori de minste kans op vertekening van het effect en kan derhalve de hoogste mate van zekerheid geven over de oorzakelijke relatie tussen de interventie en het waargenomen effect. Binnen GRADE geldt daarom dat de kwaliteit van bewijs van een goed opgezette en uitgevoerde RCT bij aanvang van de beoordeling hoog is, en die van observationele studies laag.

GRADE houdt vervolgens rekening met factoren die de kwaliteit van bewijs verlagen (*downgraden*) of verhogen (*upgraden*). Risico van bias, inconsistentie, indirect bewijs, onnauwkeurigheid en publicatiebias kunnen leiden tot downgraden. De kwaliteit van bewijs op basis van observationele studies kan hoger worden wanneer er sprake is van een (zeer) groot effect, een dosis-responsrelatie of vertekening door een ontbrekend kenmerk in tegengestelde richting waardoor het effect wordt onderschat. Daarmee kunnen goed uitgevoerde observationele studies in principe even goed of beter bewijs leveren dan RCT's met een groot risico op bias.

Bij vertaalslag van bewijs naar advies worden de in kaart gebrachte eventuele onzekerheden tezamen met contextuele factoren gewogen. Naast de kwaliteit van bewijs spelen contextuele factoren bij de beoordeling of een interventie voldoet aan SWP dus ook een rol. Het kan bijvoorbeeld gaan over wat als passend onderzoek beschouwd wordt voor de betreffende interventie en over medische aspecten.

⁸ In 2024 is ook een beoordelingskader over kosteneffectiviteit uitgebracht (Zorginstituut Nederland, 2024a). Voor de bespreking van bewijsvoering vormt het SWP-criterium het uitgangspunt. Wanneer een toetsing aan alle vier pakketcriteria aan de orde is, is ook een beoordeling van de kosteneffectiviteit nodig. Op sommige aspecten is deze richtlijn dan ook uitgebreider dan het beoordelingskader SWP. In het bijzonder bevat hij nadere handvatten om te beoordelen of de kwaliteit van bewijs dat gebaseerd is op (simulatie)modellen van voldoende kwaliteit is.

2.3 Passend bewijs volgens de Gezondheidsraad

De Gezondheidsraad is een onafhankelijke wetenschappelijke adviesraad voor regering en parlement op het brede terrein van de volksgezondheid en gezondheidszorg. Met zijn adviezen reikt de raad op wetenschap gebaseerd inzichten en handelingsopties aan en vormt zo een brug tussen wetenschap en beleid. In het kader van de richtlijn van de adviescommissie zijn vooral de adviesdomeinen vaccinatie en bevolkingsonderzoek van belang, om verschillende redenen. Voor deze terreinen geldt dat de minister het advies vraagt aan de Gezondheidsraad, en bovendien komt het besluit van de minister vrijwel altijd overeen met het advies. Het is daarom belangrijk dat het wetenschappelijk onderzoek waarop adviezen van de Gezondheidsraad op deze terreinen worden gebaseerd, langs dezelfde lat wordt gelegd als die de adviescommissie adviseert.

Vaccinaties

De Gezondheidsraad adviseert over het rijksvaccinatieprogramma en andere publieke vaccinatieprogramma's, zoals vaccinatie tegen griep. De adviezen gaan over de keuze voor vaccins, maar ook aan wie deze moeten worden aangeboden. Ze worden opgesteld door de vaste commissie Vaccinaties. De commissie hanteert daarbij een vast beoordelingskader. In het kader van de richtlijn van de adviescommissie is vooral het criterium effectiviteit van de vaccinatie van belang. Dit criterium stelt dat de vaccinatie effectief moet zijn in het voorkomen van ziekten of reduceren van symptomen.

De bewijslast die gebruikt wordt voor het beoordelen van de effectiviteit is gebaseerd op studies waar randomisatie is toegepast. Dit geldt in het bijzonder voor de werkzaamheid van een vaccin. Wanneer een vaccin dat werkzaam is gebleken opgenomen wordt in een vaccinatieprogramma, wordt de effectiviteit in de bevolking met observationele methoden vastgesteld. Het doel hiervan is de effectiviteit van het programma in de dagelijkse praktijk te monitoren. Wat uiteindelijk het effect in het niveau van de bevolking als geheel is, wordt immers ook door bijvoorbeeld het deelnamepercentage bepaald.

Bevolkingsonderzoek

Ook het adviseren over bevolkingsonderzoek is een vaste taak van de Gezondheidsraad. Deze adviezen worden opgesteld door de vaste commissie Bevolkingsonderzoek. De commissie doet dat aan de hand van internationaal gangbare criteria, de zogenaamde Wilson & Jungner-criteria. In het kader van de richtlijn van de adviescommissie is vooral het criterium 'het nut van screening moet vaststaan en opwegen tegen de risico's' van belang. Dit criterium stelt dat het moet vaststaan dat vroege opsporing van de desbetreffende ziekte(n) of aandoening(en) bij de doelgroep in kwestie kan leiden tot een significante vermindering van ziektelast, of tot andere voor de deelnemers zinvolle uitkomsten in verband met het gezondheidsprobleem waarop de screening is gericht; die voordelen moeten duidelijk opwegen tegen de nadelen die screening (voor henzelf of voor anderen) altijd ook kan hebben. Voorbeelden van dergelijke nadelen zijn de angst die een screening kan veroorzaken bij een deelnemer, en het via screening eerder ontdekken en behandelen van ziekten die anders nooit tijdens het leven tot symptomen zouden hebben geleid (overdiagnostiek en overbehandeling),

De bewijslast die gebruikt wordt voor het beoordelen van de effectiviteit van een nieuw bevolkingsonderzoek moet gebaseerd zijn op een gerandomiseerde studie. Om wijzigingen in bestaande bevolkingsonderzoeken aan te brengen (bijv. veranderingen in start leeftijd) wordt een lagere standaard gehanteerd. Hierbij worden meestal modelleringsstudies en ook observationele designs gebruikt. Wat in deze situaties als 'passend bewijs' wordt gezien, wordt van geval tot geval bekeken.

2.4 Erkende interventies

Er bestaan verschillende databanken van erkende interventies, met daarin de belangrijkste interventies uit de werkelden langdurige zorg, preventie, sport, welzijn en jeugdhulp. Een interventie kan erkenning krijgen op drie verschillende niveaus, namelijk Goed beschreven, Goed onderbouwd en Effectief. Hiertoe stellen deskundigen uit de wetenschap en praktijk de kwaliteit, uitvoerbaarheid en effectiviteit van deze interventies vast. De criteria staan beschreven in de meest recente editie van een rapport van het Samenwerkingsverband Erkenning van Interventies (2024).⁹ Erkenning is niet direct gekoppeld aan een beslissing over de inzet van publieke middelen, maar geeft beleidsmakers wel inzicht in de beschikbaarheid van werkzame interventies binnen uiteenlopende domeinen.

Voor het niveau Effectief wordt getoetst of met de interventie de gewenste verandering voor de doelgroep wordt bereikt. De bewijskracht van de effectstudie(s) bepaalt of een interventie binnen dit niveau wordt erkend op het subniveau eerste aanwijzingen, goede aanwijzingen of sterke aanwijzingen voor effectiviteit. Hiervoor hanteert het samenwerkingsverband een lichte versie van de al eerder genoemde GRADE-systematiek waarbij er binnen de erkenningscriteria ruimte is voor kwalitatief evaluatieonderzoek of een combinatie van kwalitatieve en kwantitatieve onderzoeksmethoden. Er is ook ruimte voor effectonderzoek naar complexe interventies dat gebruikmaakt van andere designs dan de RCT voor het 'best mogelijke bewijs'.

Om als een effectieve interventie op het hoogste subniveau erkend te worden moet een interventie onderbouwd zijn met ten minste twee studies (waarvan minimaal één Nederlandse studie) met sterke tot zeer sterke bewijskracht of met herhaalde casestudies met minstens tien cases, uitgevoerd in verschillende condities. Een studie heeft een zeer sterke bewijskracht als er sprake is van randomisatie of een andere opzet die de causale relatie tussen interventie en effect voldoende overtuigend aantoont. Voor een sterke bewijskracht voldoen een aantal andere designs waarvan de kern is dat er sprake is van een vergelijking met een controlegroep en er een follow-upmeting na zes maanden is gedaan.

Het beoordelingskader laat dus naast de RCT ruimte voor een breed scala aan andere evaluatiemethoden. Er is een beoordelingscommissie die beslist in welke gevallen welke evaluatiemethode passend is. Hiervoor bestaat echter geen expliciete richtlijn.

2.5 Conclusie

Deze richtlijn heeft dezelfde uitgangspunten als de werkwijzer voor MKBA in het sociaal domein en sluit op hoofdlijnen aan bij de benaderingen van ZIN en de Gezondheidsraad. De Gezondheidsraad hanteert soms een hogere bewijslast, bijvoorbeeld voor vaccinaties en bevolkingsonderzoeken. De reden hiervoor is dat het hier gaat om interventies die gezondheidsschade kunnen opleveren. Los daarvan kunnen de benaderingen van ZIN en Gezondheidsraad in sommige gevallen tot andere eisen aan passend bewijs leiden dan deze richtlijn. In de komende hoofdstukken werken we de richtlijn verder uit.

⁹ In het samenwerkingsverband werken zes kenniscentra samen: Kenniscentrum Sport & Bewegen, Movisie, Nederlands Jeugdinstituut, Rijksinstituut voor Volksgezondheid en Milieu (RIVM), Trimbos-instituut en Vilans.

3 Bewijsmethode

Deze richtlijn hanteert als uitgangspunt dat effectonderzoek ten minste gebruikmaakt van een controlegroep om te kunnen stellen dat de interventie effectief is. Gerandomiseerde experimenten, regression discontinuity designs en difference-in-differences zijn veelvoorkomende voorbeelden van methoden die hieraan voldoen.

Om in kaart te brengen of een preventiemaatregel effectief is, is het van belang om vast te stellen in hoeverre er een oorzakelijk verband is tussen de maatregel en de relevante uitkomsten. De zekerheid waarmee een onderzoek een oorzaak-gevolgrelatie kan vaststellen, wordt ook wel de interne validiteit van het onderzoek genoemd.

Voor een effectmeting met hoge interne validiteit is een overtuigende benadering van de tegenfeitelijke situatie essentieel. De tegenfeitelijke situatie geeft de uitkomsten van deelnemers aan de interventie weer wanneer zij de interventie niet zouden hebben ontvangen. Het verschil tussen de tegenfeitelijke situatie en de geobserveerde uitkomsten is het effect van de interventie. De tegenfeitelijke situatie is per definitie niet observeerbaar waardoor een onderzoeker aannames moet doen over de uitkomsten in afwezigheid van de interventie. In de meeste onderzoeken wordt hiervoor een controlegroep gebruikt.¹⁰ Als de controlegroep in alle opzichten vergelijkbaar is met de groep die de interventie heeft ontvangen, zonder dat zij zelf beïnvloed is door de interventie, kan worden aangenomen dat de uitkomsten van de controlegroep een overtuigende benadering van de tegenfeitelijke situatie zijn.

De grootste uitdaging bij studies naar effectiviteit is dat interventies vaak selectief worden ingezet. Dat wil zeggen dat de groep die een bepaalde interventie heeft ontvangen niet vergelijkbaar is met de groep die deze interventie niet heeft ontvangen. De uitkomsten van deze groep zijn dan geen overtuigende benadering van de tegenfeitelijke situatie voor de groep die de interventie heeft ontvangen. Een verschil in uitkomsten tussen beide groepen is daardoor niet (volledig) toe te wijzen aan de interventie, omdat het ook het gevolg kan zijn van de verschillen in kenmerken van de deelnemers of veranderingen buiten de context en invloed van het onderzoek. De correlatie tussen deelname en uitkomst is daarom bij selectieve inzet een onbetrouwbare schatting van het causale effect van de interventie.

Er zijn verschillende manieren om de tegenfeitelijke situatie te benaderen. Dit noemen we de bewijsmethode. Sommige bewijsmethoden leiden tot een meer overtuigende benadering van de tegenfeitelijke situatie dan andere. De Maryland Scientific Methods Scale (SMS) ordent deze methoden naar hoe overtuigend ze het probleem van selectiviteit adresseren en daarmee naar hoe overtuigend ze de tegenfeitelijke situatie benaderen (zie Box 3.1). Hierbij merken we op dat er naast de bewijsmethode ook andere aspecten van het onderzoeksdesign zijn die de interne validiteit beïnvloeden. Denk aan de grootte van de steekproef, de (objectieve) meetbaarheid en relevantie van de gehanteerde uitkomstmaten en controlevariabelen of eventuele clustering van de standaardfouten.¹¹ Deze aspecten staan echter niet centraal in dit hoofdstuk.

¹⁰ In de medische literatuur wordt dit ook wel de comparator genoemd en is dit onderdeel van de PICOT.

¹¹ Het verdient aanbeveling om bij de opzet van nieuw onderzoek over deze aspecten van het onderzoeksdesign na te denken, samen met de bewijsmethode. Het komt dan ook steeds vaker voor dat deze keuzes in het design vooraf worden vastgelegd in de vorm van een preregistratie, waardoor de validiteit van nieuw onderzoek als het ware al voor de start van het onderzoek getoetst kan worden.

Box 3.1 De Maryland Scientific Methods Scale

De Maryland Scientific Methods Scale (SMS) onderscheidt vijf niveaus in de bewijskracht van kwantitatieve bewijsmethoden:

5. Gerandomiseerd experiment
Gerandomiseerde onderzoeksoopzetten waarbij expliciete randomisatie plaatsvindt in behandel- en controlegroepen. Er wordt uitgebreid bewijs geleverd voor de vergelijkbaarheid van beide groepen waarbij geen significante verschillen in niveaus of trends mogen bestaan tussen de groepen. Controlevariabelen kunnen worden gebruikt, maar mogen de hoofdresultaten nauwelijks beïnvloeden. Er is aandacht voor selectieve uitval en voor mogelijke beïnvloeding van de controlegroep die beperkt of idealiter afwezig is.
4. Quasi-experimentele analyse of natuurlijke experimenten
Analyse maakt gebruik van (zo goed als) willekeurige toewijzing van de interventie die ontstaat als gevolg van bestaande instituties of regelgeving. Vaak wordt hierbij een instrument of een discontinuïteit gebruikt waarvan de geschiktheid overtuigend moet worden aangetoond en onderbouwd. Controlevariabelen kunnen worden gebruikt, maar mogen de hoofdresultaten nauwelijks beïnvloeden.
3. *Difference-in-differences* of andere panelmethoden¹²
Vergelijking van uitkomsten in de interventiegroep met een geloofwaardige controlegroep, zowel voor als na de interventie. De keuze voor de controlegroep wordt gemotiveerd en er wordt bewijs gegeven voor de vergelijkbaarheid met de behandelde groep. Controlevariabelen of matchingtechnieken kunnen worden toegepast om verschillen tussen groepen te corrigeren.
2. Cross-sectionele analyse of voor-navergelijking met passende controlevariabelen
In een cross-sectionele analyse worden controlevariabelen of matchingtechnieken toegepast om verschillen tussen groepen te corrigeren. In een voor-navergelijking worden controlevariabelen gebruikt om veranderingen tussen de twee perioden te corrigeren.
1. Cross-sectionele analyse of voor-navergelijking zonder controlevariabelen
Een cross-sectionele vergelijking tussen behandelde en onbehandelde personen óf een voor-navergelijking binnen de behandelde groep, zonder onbehandelde controlegroep. Er worden geen controlevariabelen gebruikt om verschillen tussen groepen of perioden statistisch te corrigeren.

Bron: What Works Centre for Local Economic Growth (2016)

Het uitgangspunt van deze richtlijn is dat een interventie ten minste op niveau 3 van de Scientific Methods Scale (SMS) onderzocht moet zijn om te kunnen stellen dat de interventie effectief is. Dit betekent dat er een bewijsmethode is gebruikt met een goede controlegroep waardoor niet alleen veranderingen in uitkomsten maar daadwerkelijke effecten van een interventie in kaart worden gebracht. Systematische reviews scoren we aan de hand van de onderliggende studies. Een systematische review die alleen maar kijkt naar gerandomiseerde experimenten scoort dan bijvoorbeeld niveau 5. Bewijsmethoden van niveau 3 of hoger waarborgen een minimale kwaliteit

¹² Panelmethoden zijn gebaseerd op data waarin dezelfde individuen door de tijd worden gevolgd, zodat kan worden gecontroleerd voor kenmerken die constant blijven. Er bestaan verschillende panelmethoden die onder niveau 3 van de Maryland Scale vallen, zoals *fixed effects* of *first differences*-modellen.

van bewijs die nodig is om betrouwbare uitspraken over effecten te kunnen doen. Zonder een goede controlegroep is het risico namelijk groot dat veranderingen in de uitkomstvariabele niet het gevolg zijn van de interventie maar van andere factoren. Dit uitgangspunt sluit aan bij de werkwijzer voor MKBA in het sociaal domein.

Voor specifieke interventies kan een hoger bewijsniveau dan niveau 3 wenselijk zijn. Bijvoorbeeld als de preventieve interventie ook tot gezondheidsschade kan leiden, zoals bij vaccinaties of bij bevolkingsonderzoeken. Vaccins kunnen verschillende negatieve bijwerkingen hebben¹³ en na een (vals)positieve uitslag in een preventieve screening op darmkanker moet er bijvoorbeeld een coloscopie gedaan worden met een risico op ernstige complicaties. Hierdoor is meer zekerheid over de positieve gezondheidseffecten wenselijk voordat dergelijke interventies grootschalig worden ingezet. Bij veel preventiemaatregelen is er echter geen risico op gezondheidsschade en zal niveau 3 volstaan. Een nuttig raamwerk voor het bepalen of een hoger bewijsniveau nodig is, is de Value of Information-analyse.¹⁴ Dit is een hulpmiddel om te bepalen hoeveel het waard is om extra onzekerheid weg te nemen dat gebruikt kan worden om de kosten van extra bewijs af te wegen tegen de baten.

Het vervolg van dit hoofdstuk bespreekt enkele veelgebruikte bewijsmethoden onder de verschillende niveaus van de SMS. Hierbij baseren we ons op een uitwerking van de SMS door het What Works Centre for Local Economic Growth (2016). Waar relevant verwijzen we bovendien door naar uitgebreidere discussies van de specifieke bewijsmethoden. Iedere methode wordt toegelicht aan de hand van voorbeelden van toepassingen op preventieonderzoek.¹⁵ Deze voorbeelden worden uitgebreider besproken in Bijlage B.

3.1 SMS-niveau 5

Het hoogste niveau van de Maryland Scale bestaat uit gerandomiseerde experimenten (*Randomized Controlled Trials*). In een gerandomiseerd experiment worden deelnemers willekeurig toegewezen aan de interventie- of controlegroep. Door deze willekeurige toewijzing zijn deelnemers in de interventiegroep vergelijkbaar met deelnemers in de controlegroep, op het ontvangen van de interventie na. Hierdoor is er geen selectiviteit en zijn de uitkomsten van de controlegroep dus een overtuigende benadering van de tegenfeitelijke situatie voor die van de interventiegroep. Om de effectiviteit van de interventie te bepalen kunnen dan simpelweg de uitkomsten van deelnemers in de interventiegroep vergeleken worden met die van deelnemers in de controlegroep.¹⁶

Een goed gerandomiseerd experiment voldoet aan enkele voorwaarden:

- **Succesvolle randomisering.** De toewijzing aan de interventie- en controlegroep is volledig willekeurig. Dit kan getoetst worden door de kenmerken van deelnemers in de interventie- en controlegroep te vergelijken. Als er toch significante verschillen zijn tussen de kenmerken van de deelnemers in de twee groepen, dan is er sprake van selectiviteit. Dan zijn de uitkomsten van de controlegroep geen overtuigende benadering meer van de tegenfeitelijke

¹³ <https://www.lareb.nl/vaccin-kennis>

¹⁴ Zie Zorginstituut Nederland (2024b) voor meer informatie.

¹⁵ Voor meer voorbeelden van bewijsmethoden en toepassingen op leefstijlinterventies verwijzen we naar de Handreiking bewijsmethoden leefstijlinterventies in de zorg van de Coalitie Leefstijl in de Zorg (2025).

¹⁶ Er bestaan verschillende varianten op de standaard-RCT, zoals hier beschreven. Voor een uitgebreid overzicht, zie Kim et al. (2021).

situatie en kunnen de verschillen in uitkomsten niet meer (volledig) aan de interventie worden toegewezen.¹⁷

- **Geen beïnvloeding van de controlegroep.** Deelnemers uit de controlegroep mogen niet worden blootgesteld aan de interventie. Blootstelling kan bijvoorbeeld gebeuren wanneer deelnemers uit de interventiegroep informatie of middelen delen met de controlegroep.
- **Geen selectieve uitval of gecontroleerd voor selectieve uitval.** Selectieve uitval treedt op wanneer deelnemers gedurende het experiment uitvallen en de kans daarop verschilt tussen de interventie- en de controlegroep. Hierdoor wordt de oorspronkelijke willekeurige verdeling verstoord en zijn de interventie- en de controlegroep niet meer goed vergelijkbaar. Als er toch selectieve uitval plaatsvindt gedurende een experiment, dan is het in sommige gevallen mogelijk om te controleren voor selectieve uitval, als er duidelijke factoren zijn die dit verklaren. Deze factoren kunnen dan als controlevariabelen worden meegenomen in de analyse om de interventie- en de controlegroep weer goed vergelijkbaar te maken.

Dubbele blindering in een gerandomiseerd experiment heeft de voorkeur om bias als gevolg van de gecontroleerde experimentele setting te reduceren en de kans op succesvolle randomisatie te vergroten. Dubbele blindering betekent dat zowel deelnemers als onderzoekers niet weten wie in de behandel- en wie in de controlegroep zit. Dit voorkomt dat verwachtingen of voorkennis het gedrag van deelnemers of de interpretatie van resultaten beïnvloeden waardoor bias wordt geminimaliseerd en de betrouwbaarheid van de uitkomsten toeneemt. Dubbele blindering draagt daarnaast bij aan de kans op succesvolle randomisatie, omdat kennis van de verdeling over de interventie- en controlegroepen er anders toe kan leiden dat deelnemers de interventie weigeren of dat onderzoekers toch een andere interventie toedienen.¹⁸ Dubbele blindering is echter niet bij elke interventie mogelijk, zeker in het geval van preventie.

Het belangrijkste voordeel van een gerandomiseerd experiment is de hoge interne validiteit die behaald kan worden door randomisatie. Indien het experiment voldoet aan de bovengenoemde voorwaarden geeft het de meest zuivere schatting van het effect van een interventie. Hiertegenover staan echter ook verschillende nadelen. Een experiment gaat vaak gepaard met hoge uitvoeringskosten, soms met ethische bezwaren¹⁹ en kan een beperkte externe validiteit hebben vanwege de gecontroleerde experimentele setting en selectieve deelname aan het experiment.²⁰ In sommige gevallen kan het doen van observationeel onderzoek daarom passender zijn.

Veelvoorkomende varianten van de RCT zijn de *stepped wedge*-RCT en de cluster-RCT. In de *stepped wedge*-RCT wordt het moment van de start van de interventie gerandomiseerd waardoor sommige

¹⁷ Door te controleren voor observeerbare verschillen in karakteristieken kan hiervoor gedeeltelijk worden gecorrigeerd, maar het valt niet uit te sluiten dat er nog andere niet-observeerbare verschillen zijn die de schatting van het effect beïnvloeden.

¹⁸ Voor meer informatie, zie de Cochrane Risk of Bias Tool (Sterne et al., 2019).

¹⁹ Zo kan het bijvoorbeeld als onethisch worden gezien om een deelgroep (bijvoorbeeld kinderen) de toegang tot een mogelijk waardevolle interventie te weigeren. In dat geval kunnen alternatieve varianten van de basis-RCT helpen, omdat bijvoorbeeld in een *stepped wedge*-RCT alleen het moment van de start van de interventie gerandomiseerd wordt en uiteindelijk wel alle deelnemers toegang krijgen tot de interventie.

²⁰ Met betrekking tot de externe validiteit is de werving en selectie van deelnemers een punt van aandacht. Dit is specifiek uitdagend bij interventies die gericht zijn op een lastig te bereiken doelgroep, zoals in het experiment met CenteringZwangerschap (Wagij et al., 2024). Deze interventie richtte zich met name op zwangere vrouwen met een kwetsbare achtergrond. Om de representativiteit van deze groep in het experiment te vergroten werd daarom gebruikgemaakt van verschillende lokale netwerken van moeders om deelnemers te werven.

deelnemers de interventie eerder krijgen dan anderen. Op verschillende meetmomenten kunnen de uitkomsten van deelnemers die de interventie al hebben gekregen dan worden vergeleken met die van deelnemers die de interventie nog niet hebben gekregen. Het voordeel van de *stepped wedge*-RCT is dat geen enkele deelnemer een mogelijk waardevolle interventie wordt geweigerd.²¹ In een cluster-RCT worden geen individuen, maar groepen individuen, toegewezen aan de interventie- of controlegroep. Een cluster kan bijvoorbeeld een school, ziekenhuis, wijk of huisartspraktijk zijn. Een cluster-RCT kan worden gebruikt wanneer een interventie alleen op groepsniveau kan worden toegediend (Gail et al., 1996) of wanneer er een risico is op beïnvloeding van de controlegroep (Torgerson, 2001). Door groepen die veel met elkaar in contact komen tegelijk te randomiseren wordt het risico kleiner dat deelnemers uit de controlegroep toch met de interventie in aanraking komen.

Er zijn verschillende voorbeelden van RCT's in het preventiedomein. Zo onderzoeken De Koning et al. (2020) het effect van preventieve longkankerscreening waarbij deelnemers willekeurig zijn toegewezen aan een interventiegroep die screening ontvangt of een controlegroep waarvoor dat niet zo is. Een ander voorbeeld is een experiment met een systeeminterventie voor obesitas onder kinderen waarin een *stepped wedge* cluster-RCT wordt toegepast (Allender et al., 2021). In dit onderzoek worden gehele gemeenschappen willekeurig in de interventie- of controlegroep ingedeeld. In gemeenschappen in de interventiegroep start de interventie vervolgens in 2015 en in gemeenschappen in de controlegroep pas in 2019. Tussen 2015 en 2019 zijn de uitkomsten van inwoners in de interventiegemeenschappen vergeleken met die van inwoners in de controlegemeenschappen om effecten vast te stellen.

3.2 SMS-niveau 4

Op niveau vier van de Maryland Scale staan natuurlijke experimenten. In een natuurlijk experiment wordt variatie gebruikt in de toewijzing van een interventie die vaak ontstaat als gevolg van bestaande instituties of regelgeving. Deze variatie zorgt ervoor dat van twee vergelijkbare personen er slechts één meedoet aan de interventie. Denk bijvoorbeeld aan beleidswijzigingen, geografische grenzen of lotingen die buiten de invloed van de onderzoeker om bepalen wie wel of niet wordt blootgesteld aan een interventie. Hieronder bespreken we twee veelgebruikte bewijsmethoden bij natuurlijke experimenten: *regression discontinuity design* en instrumentele variabelen.

Regression discontinuity design

De eerste methode die we bespreken is een *regression discontinuity design* (RDD).²² Een RDD kan worden toegepast als de toewijzing aan een interventie (direct of indirect) afhankelijk is van een bepaalde drempelwaarde, zoals een maximaal inkomensniveau of een minimale leeftijd. Personen net boven en net onder deze drempel worden met elkaar vergeleken, omdat zij in vrijwel alle opzichten vergelijkbaar zijn, behalve dat de ene groep de interventie krijgt en de andere niet.

In het preventiedomein wordt in verschillende onderzoeken gebruikgemaakt van RDD's, zoals in een onderzoek naar de effectiviteit van leefstijladvies voor mensen met diabetes (Lemp et al., 2023). Het onderzoek maakt gebruik van een drempelwaarde in de bloedsuikerspiegel die bepaalt wie er wordt doorverwezen naar het leefstijladvies. Vervolgens worden mensen met een bloedsuikerspiegel net boven de drempelwaarde van 42 mmol/mol vergeleken met mensen met een bloedsuikerspiegel net onder deze drempelwaarde.

²¹ Zie Hemming et al. (2014) voor meer informatie.

²² Voor meer informatie over RDD's, zie ook Hoofdstuk 6 van Angrist en Pischke (2008).

Een overtuigende RDD voldoet aan de volgende voorwaarden:

- **Daadwerkelijke discontinuïteit op de drempelwaarde.** De drempel voor de interventie moet duidelijk en daadwerkelijk toegepast worden, zodat bij een vergelijking van individuen boven en onder de drempelwaarde daadwerkelijk personen die de interventie hebben ontvangen worden vergeleken met mensen die dat niet hebben ontvangen.
- **Discontinuïteit heeft op geen andere manier invloed op de uitkomst.** Op de drempelwaarde mag alleen het wel of niet in aanmerking komen voor de interventie veranderen. Anders kan niet meer het volledige effect aan de interventie worden toegewezen. Dit betekent onder andere dat het niet mogelijk is om invloed uit te oefenen op het wel of niet halen van de drempelwaarde.²³ Indien dat wel mogelijk is, is de toewijzing aan de interventie niet meer willekeurig en zijn de uitkomsten van de controlegroep geen overtuigende benadering meer van de tegenfeitelijke situatie voor de uitkomsten van de interventiegroep. Ook mogen er geen andere discontinuïteiten spelen op de drempelwaarde.²⁴ Wie 18 wordt, mag bijvoorbeeld legaal alcohol drinken en daarom lijkt dit op het eerste oog mogelijk een goede discontinuïteit om het gezondheidseffect van alcoholgebruik te onderzoeken. Deze leeftijdsgrens wordt echter ook gebruikt voor allerlei andere zaken (zoals het hebben van een eigen risico in de zorgverzekering, het halen van een rijbewijs) die mogelijk ook de gezondheid beïnvloeden waardoor het geschatte effect niet volledig aan alcoholgebruik kan worden toegewezen.

Een nadeel van een RDD is dat het in principe alleen een lokaal effect van de interventie identificeert, wat de toepasbaarheid van de uitkomsten beperkt.²⁵ Het effect van een interventie wordt namelijk alleen in kaart gebracht voor personen rondom de drempelwaarde en het is niet altijd aannemelijk dat eenzelfde effect verwacht kan worden voor groepen die veel verder van de drempelwaarde af zitten.

Er zijn verschillende varianten van de standaard-RDD beschikbaar. Dit zijn vaak combinaties tussen een RDD en een andere bewijsmethode. Zo kan een RDD gecombineerd worden met een instrumentele variabele (zie het volgende kopje 'Instrumentele variabelen') als de drempelwaarde in de RDD niet volledig bepaalt wie de interventie krijgt, maar alleen de kans op het ontvangen van de interventie verhoogt. Deze methode wordt ook wel een *fuzzy RDD* genoemd. Een RDD kan ook gecombineerd worden met *difference-in-differences* (zie SMS-niveau 3) om het effect van de interventie te onderscheiden van andere discontinuïteiten op de drempelwaarde. Onderzoek naar de effecten van een hogere suikertaks in Noorwegen past deze methode bijvoorbeeld toe (Øvrebo et al., 2020). Hierbij wordt gebruikgemaakt van een drempelwaarde in de tijd, omdat op 1 januari 2018 de accijns op snoep en frisdrank wordt verhoogd. Een reguliere RDD, waarbij de verkoop van snoep en frisdrank net voor 1 januari wordt vergeleken met die van net na 1 januari, zou een vertekend beeld

²³ In het geval dat er een risico bestaat op manipulatie kan worden gekeken of er een piek is in het aantal personen met een waarde net onder of boven de drempelwaarde. Dergelijke pieken zijn een signaal van manipulatie. Manipulatie rondom de drempelwaarde is een reëel risico, zo blijkt bijvoorbeeld uit onderzoek naar de effecten van de Vogelwaardwijken, een subsidie om de leefbaarheid in Nederlandse achterstandswijken te verbeteren (Gerritsen et al., 2017). Hieruit bleek dat er selectieve *non-compliance* was van de toewijzingsregel.

²⁴ Dit kan getoetst worden door personen rondom de grenswaarde te vergelijken op waargenomen kenmerken. Zij mogen niet significant van elkaar verschillen.

²⁵ Met additionele aannames is het echter wel mogelijk om ook uitspraken te doen over effecten voor groepen verder van de drempelwaarde, zie Angrist en Rokkanen (2015).

van de effectiviteit van de interventie geven. Het seizoenseffect, waarbij de verkoop van snoep en frisdrank standaard lager is in de eerste maanden van een jaar dan in de laatste maanden, zou dan namelijk ook aan de interventie worden toegewezen. Om dit seizoenseffect te onderscheiden van het effect van de interventie vergelijkt het onderzoek de verandering in de verkoop van snoep en frisdrank voor en na 1 januari 2018 met de verandering voor en na 1 januari 2017.

Instrumentele variabelen

De tweede methode die onder niveau 4 van de Maryland Scale valt is een instrumentele variabelen (IV) analyse.²⁶ Bij een IV-analyse worden één of meerdere variabelen gebruikt die wel van invloed zijn op de toewijzing aan de interventie- of controlegroep, maar die geen direct effect hebben op gezondheidsuitkomsten. Een voorbeeld van een instrumentele variabele is variatie in de voorkeuren van artsen voor bepaalde behandelingen. Dit principe is toegepast in een onderzoek naar de mentale gezondheidseffecten van medicatie die helpt bij het stoppen met roken (Thomas et al., 2013). Artsen verschillen namelijk in hun voorschrijfgedrag: een vergelijkbare patiënt krijgt bij de ene arts wél een bepaald medicijn voorgeschreven, terwijl een andere arts dit medicijn niet voorschrijft. Deze variatie in voorschrijfgedrag kan worden ingezet als instrumentele variabele om het effect van de medicatie te onderzoeken.

Een goede instrumentele variabele voldoet aan de volgende:

- **Relevantie.** Het instrument moet een voldoende groot deel van de variatie in het wel of niet ontvangen van de interventie verklaren.²⁷ Als de relatie tussen het instrument en het krijgen van de interventie zwak is of erg onnauwkeurig gemeten, dan kan er niet genoeg informatie uitgehaald worden over het effect van de behandeling.
- **Exclusie.** Het instrument mag de onderzochte uitkomsten alléén beïnvloeden via de interventie, en niet via een andere route. In het onderzoek naar de effectiviteit van medicatie die helpt bij het stoppen met roken, waarin variatie in de voorkeuren van artsen voor het voorschrijven van medicatie wordt gebruikt als instrument, betekent dit het volgende. Artsen met een voorkeur voor het voorschrijven van de stoppen-met-rokenmedicatie mogen daarnaast niet ook vaker een ander medicijn voorschrijven dat eveneens effect heeft op mentale gezondheidsuitkomsten. In dat geval is het namelijk niet meer duidelijk of het verschil in gezondheidseffecten het gevolg is van de stoppen-met-rokenmedicatie of van het andere medicijn.
- **Exogeniteit.** Het instrument mag niet samenhangen met (niet-geobserveerde) factoren die ook de uitkomst beïnvloeden. Dat betekent dat in het bovengenoemde voorbeeld de voorkeur van de arts voor de stoppen-met-rokenmedicatie niet mag worden veroorzaakt door (niet-geobserveerde) verschillen tussen de patiënten die de verschillende artsen behandelen. In dat geval kunnen verschillen in gezondheidsuitkomsten namelijk ook veroorzaakt worden door de verschillen in patiënten in plaats van door het gebruik van de stoppen-met-rokenmedicatie.
- **Monotoniciteit.** Het instrument beïnvloedt alle individuen in dezelfde richting wat betreft hun kans om de interventie te ontvangen. Dat betekent: als het instrument de kans op de

²⁶ Er bestaan verschillende varianten op de standaard instrumentele variabele analyse, zoals hier beschreven. Voor een uitgebreid overzicht, zie Walker et al. (2024). Voor meer informatie over instrumentele variabelen, zie ook Hoofdstuk 4 van Angrist en Pischke (2008).

²⁷ De relevantie van het instrument wordt typisch getoetst in een regressie van de interventie op het instrument. Hierbij geldt als vuistregel dat als de F-statistiek van deze regressie ten minste de waarde van 10 aanneemt, dan is het instrument relevant (Staiger & Stock, 1997).

interventie vergroot voor sommige personen, dan mag het die kans voor niemand verkleinen. In het voorbeeld houdt dit in dat er geen patiënten zijn die de stoppen-met-rokenmedicatie wél krijgen voorgeschreven door een arts met een geringe voorkeur, maar níet van een arts die juist een sterke voorkeur voor deze middelen heeft (en omgekeerd). Met andere woorden: behandeling door een arts met een voorkeur voor de stoppen-met-rokenmedicatie mag voor geen enkele patiënt de kans verkleinen om deze medicijnen voorgeschreven te krijgen.

Een nadeel van een IV-schatting is dat deze methode alleen een lokaal effect van de interventie identificeert. Met instrumentele variabelen kan namelijk alleen het effect van de interventie in kaart worden gebracht voor personen die door het instrument de behandeling wel of niet krijgen. In het bovengenoemde voorbeeld gaat het om iemand die wel de stoppen-met-rokenmedicatie zou krijgen voorgeschreven door een arts met een sterke voorkeur, maar niet bij een arts met een minder sterke voorkeur. Oftewel, de analyse brengt niet de effecten in kaart voor patiënten die door iedere arts deze medicatie voorgeschreven zouden krijgen of voor patiënten die door geen enkele arts deze middelen voorgeschreven zouden krijgen. Dit beperkt de toepasbaarheid van deze methodiek. Het effect van een interventie op de hele populatie hoeft namelijk niet overeen te komen met het effect voor deze specifieke groep.

3.3 SMS-niveau 3

Niveau 3 op de Maryland Scale bestaat uit methoden die de uitkomsten van de interventiegroep zowel voor als na de invoering van de interventie vergelijken met de uitkomsten van een geloofwaardige controlegroep. Net als bij niveau 4 wordt er gebruikgemaakt van variatie in de toewijzing tot de interventiegroep die bijvoorbeeld ontstaat als gevolg van instituties of regelgeving. Deze methoden corrigeren voor andere factoren (zoals tijdstrends, beleidsschokken of toevallige gebeurtenissen) die de uitkomsten van de interventie- en controlegroep allebei over tijd beïnvloeden en waardoor een voor-navergelijking geen betrouwbare effectmeting oplevert (zie SMS-niveau 1 en 2). Ook corrigeren deze methoden voor verschillen in kenmerken tussen de interventie- en controlegroep, voor zover deze over tijd constant blijven. Een voorbeeld van een veelgebruikte methode is de *difference-in-differences*-analyse. Deze wordt hieronder toegelicht.

Difference-in-differences

In *difference-in-differences* worden de uitkomsten van de interventie- en controlegroep voor en na invoering van de interventie met elkaar vergeleken.²⁸ Als het verschil in uitkomsten tussen de twee groepen over tijd is veranderd, wordt dit verschil toegewezen aan de interventie. De toewijzing aan de interventie- of controlegroep kan daarbij worden veroorzaakt door een beleidsregel. Onderzoek naar de effecten van een verlaging van de minimumleeftijd voor alcoholconsumptie in de Verenigde Staten maakt bijvoorbeeld gebruik van het feit dat een verlaging van de minimumleeftijd alleen voor enkele specifieke leeftijdsgroepen de toegankelijkheid van alcohol beïnvloedt (Carpenter & Dobkin, 2011). De leeftijdsverlaging heeft uitsluitend effect voor jongeren die zich bevinden tussen de oude en de nieuwe minimumleeftijd. Daarom worden in deze studie de uitkomsten van deze groep vergeleken met die van jongeren die ouder zijn dan de oorspronkelijke minimumleeftijd en met die van jongeren die jonger zijn dan de nieuwe minimumleeftijd.

De belangrijkste aanname van *difference-in-differences* is dat de geobserveerde verandering in uitkomsten van de controlegroep ook zou hebben plaatsgevonden in de interventiegroep, in het geval zij niet de interventie zouden hebben gekregen. Dit heet de 'parallelle-trends aanname'. Deze

²⁸ Voor meer informatie over *difference-in-differences*, zie Hoofdstuk 5 van Angrist en Pischke (2008).

aanname kan niet direct worden getoetst, omdat we de uitkomsten van de interventiegroep in afwezigheid van de interventie niet observeren. Deze aanname kan wel aannemelijk gemaakt worden door te kijken naar de ontwikkelingen in uitkomsten van de interventie- en controlegroep vóór invoering van de interventie. Hiervoor worden de uitkomsten van beide groepen op ten minste twee momenten voor de start van de interventie met elkaar vergeleken. Als het verschil tussen deze uitkomsten voor de start van de interventie niet (significant) verandert – oftewel de uitkomsten van de twee groepen ontwikkelden zich voor de invoering parallel aan elkaar – dan is het aannemelijk dat dit ook zou hebben doorgezet als de interventie niet zou zijn ingevoerd. Er blijft echter een risico op selectiviteit bestaan.

Een overtuigende *difference-in-differences*-analyse voldoet daarom aan de volgende voorwaarden:

- **Geloofwaardige controlegroep.** Op basis van parallelle trends voor de interventie, een onderbouwing van een beperkt risico op selectiviteit en een vergelijking van de kenmerken van de interventie- en controlegroep wordt aannemelijk gemaakt dat de controlegroep de tegenfeitelijke situatie van de behandelgroep goed benadert.
- **Geen beïnvloeding van de controlegroep.** Deelnemers uit de controlegroep mogen niet (indirect) worden blootgesteld aan de interventie. Blootstelling kan bijvoorbeeld plaatsvinden wanneer deelnemers over tijd wisselen tussen de interventie- en controlegroep of deelnemers uit de interventiegroep informatie of middelen delen met de controlegroep.

Difference-in-differences wordt regelmatig met andere methoden gecombineerd om de geloofwaardigheid van de controlegroep te verhogen. Zo wordt vaak gebruikgemaakt van matchingmethoden of *Inverse Probability Weighting* om de vergelijkbaarheid van de controlegroep met de interventiegroep te vergroten, zoals in het onderzoek naar de effecten van lokale proactieve buurtteams op onnodig gebruik van ziekenhuiszorg (Sevak et al., 2018). Tegelijkertijd kan matching het risico op selectiviteit in *difference-in-differences* niet geheel wegnemen. Dit blijkt onder andere uit de resultaten van een later uitgevoerde RCT naar dezelfde interventie (Finkelstein et al., 2020). Waar Sevak et al. op basis van *difference-in-differences* concludeerden dat de inzet van proactieve zorgteams leidde tot minder onnodig zorggebruik, kwam uit deze RCT naar voren dat de interventie waarschijnlijk niet effectief was. Dit voorbeeld illustreert dat de keuze voor een bewijsmethode op een lager niveau van de Maryland Scale gepaard gaat met een groter risico op vertekening en dat dit in de praktijk kan leiden tot een onterechte conclusie over de effectiviteit van een interventie. Daarom is het belangrijk om de voorwaarden voor een overtuigende toepassing kritisch te beoordelen.

Een veelvoorkomende variant van *difference-in-differences* is de *staggered difference-in-differences* waarbij net als in de *stepped wedge*-RCT het *stepped wedge design* wordt toegepast.²⁹ Hierbij wordt een interventie op verschillende momenten ingevoerd voor verschillende groepen waardoor groepen waar de interventie later start kunnen worden gebruikt als controlegroep voor groepen waarbij de interventie eerder is gestart. Deze methode wordt bijvoorbeeld toegepast in het onderzoek van Margaryan (2021) naar de gezondheidseffecten van de invoering van milieuzones in Duitsland. Omdat de milieuzones gefaseerd zijn ingevoerd, konden de gezondheidsuitkomsten van inwoners in steden waar al wel een milieuzone was ingevoerd, worden vergeleken met die van inwoners in steden waar dat nog niet het geval was.

²⁹ Er zijn daarnaast ook nog andere varianten van *difference-in-differences*. Zie Baker et al. (2025) voor een uitgebreid overzicht.

De *difference-in-differences*-methode is de afgelopen jaren sterk doorontwikkeld. Waar de klassieke *difference-in-differences* gebaseerd was op twee groepen die op twee momenten in de tijd met elkaar vergeleken werden, zijn inmiddels diverse varianten ontwikkeld die het mogelijk maken effecten te schatten op basis van meerdere groepen en meetmomenten (waaronder de *staggered difference-in-differences*). Er is daarnaast veel aandacht besteed aan de gevolgen van heterogeniteit van effecten tussen groepen of over de tijd, en aan oplossingen voor het probleem van niet-parallele trends. Een uitgebreid overzicht van deze recente ontwikkelingen is te vinden in Roth et al. (2023).

3.4 SMS-niveaus 1 en 2

Niveaus 1 en 2 op de Maryland Scale bestaan uit cross-sectionele analyses en voor-navergelijkingen. In een cross-sectionele analyse worden ontvangers van een interventie vergeleken met niet-ontvangers, terwijl in een voor-navergelijking de ontvangers worden vergeleken met hun eigen situatie vóór de interventie. Bij beide methoden kan niet worden uitgesloten dat het geschatte effect wordt veroorzaakt door andere factoren dan de interventie. Wordt in de analyse gecorrigeerd voor observeerbare verschillen in kenmerken tussen ontvangers en niet-ontvangers, of wordt de controlegroep samengesteld via matching, dan betreft het niveau 2; zonder dergelijke correcties gaat het om niveau 1.

Cross-sectionele analyse

Een cross-sectionele analyse is een (niet-gerandomiseerde) onderzoeksopzet waarbij data op één enkel moment in de tijd worden verzameld en geanalyseerd. Op dit ene moment worden vervolgens ontvangers van een interventie vergeleken met mensen die de interventie niet hebben ontvangen. Controlevariabelen of matchingtechnieken kunnen worden toegepast om voor verschillen tussen groepen te corrigeren.

In een cross-sectionele analyse kan niet goed worden gecontroleerd voor selectiviteit van de interventie. Wanneer deelname aan een interventie op vrijwillige basis of op voorschrift van een arts plaatsvindt, verschillen de ontvangers van de interventie waarschijnlijk in kenmerken van de niet-ontvangers. In het eerste geval zijn zij bijvoorbeeld gemotiveerder of verwachten zij te profiteren van de interventie. In dat geval leidt een cross-sectionele analyse waarschijnlijk tot een overschatting van het effect van de interventie. In het tweede geval zijn ontvangers van de interventie mogelijk minder gezond dan de niet-ontvangers. Hierdoor zal een cross-sectionele analyse het effect van de interventie onderschatten. In het meest extreme geval kan uit een dergelijke analyse zelfs komen dat de interventie een negatief effect heeft op de gezondheid, terwijl de interventie in feite positief bijdraagt aan de gezondheid.

Het risico op bias in een cross-sectionele analyse wordt duidelijk geïllustreerd in een onderzoek naar *Workplace Wellness*-programma's (Song & Baicker, 2019). Verschillende eerdere onderzoeken vonden op basis van een cross-sectionele analyse significante positieve effecten van *Workplace Wellness*-programma's (zie Pelletier, 2011 voor een overzicht). Maar de vrijwillige deelname aan deze programma's zorgt voor selectiviteit: gezonde mensen doen vaker mee aan dit soort programma's. Hierdoor kan niet het volledige gevonden effect aan de interventie worden toegeschreven. Song en Baicker (2019) vinden vervolgens in een gerandomiseerd experiment geen significante effecten van het programma, wat aangeeft dat de eerder gevonden positieve effecten waarschijnlijk bepaald werden door de selectieve deelname aan het programma.

Voor-navergelijking

In een voor-en-na vergelijking of *interrupted time-series* wordt het effect van een interventie geschat door de situatie voor en na invoering van de interventie met elkaar te vergelijken. Hierbij worden de uitkomsten van de interventiegroep nadat ze de interventie hebben gekregen vergeleken met hun uitkomsten voordat ze de interventie kregen. Er wordt dus geen gebruik gemaakt van een controlegroep. Controlevariabelen kunnen worden toegevoegd om voor veranderingen tussen de twee perioden te corrigeren.

Een voor-navergelijking corrigeert niet voor de mogelijkheid dat andere factoren (zoals trends, beleidsschokken of toevallige gebeurtenissen) de verandering in uitkomsten hebben veroorzaakt. Ook als de interventiegroep de interventie niet zou hebben ontvangen, dan zouden deze externe ontwikkelingen alsnog hebben plaatsgevonden. Denk bijvoorbeeld aan het effect van een interventie die mensen helpt met afvallen. Als deze interventie vooral gebruikt wordt door mensen die al gemotiveerd waren om af te vallen, dan is het niet aannemelijk dat de volledige daling in gewicht het gevolg is van de interventie. Gebruikers van de interventie zouden waarschijnlijk ook al zijn afgevallen zonder de interventie. Hierbij geldt dat hoe langer de tijdsperiode tussen de twee meetmomenten, hoe groter het risico dat andere factoren de ontwikkelingen in uitkomsten beïnvloeden. Dit speelt in het bijzonder bij preventieonderzoek, omdat de relevante effecten zich vaak pas op de langere termijn manifesteren.

De beperkingen van een voor-navergelijking komen duidelijk naar voren in onderzoek naar het verminderen van zorggebruik bij mensen met extreem hoge zorgkosten (Finkelstein et al., 2020): veel van deze mensen hebben in de toekomst vanzelf lagere kosten, ook zonder interventie. Dit is een veelvoorkomend fenomeen waarbij extreme waarden na verloop van tijd vanzelf terug bewegen richting het gemiddelde. Als onderzoekers alleen voor en na een interventie meten, kan deze daling in zorgkosten ten onrechte worden toegeschreven aan de interventie.

Ook het onderzoek van Bartelink et al. (2019) naar de effecten van de Gezonde Basisschool van de Toekomst illustreert de beperkingen van een voor-navergelijking. In dit onderzoek werd met een *difference-in-differences*-analyse (SMS-niveau 3) de Body Mass Index (BMI) van de interventiegroep over de tijd vergeleken met die van een controlegroep. De resultaten laten zien dat de BMI van de interventiegroep daalt ten opzichte van de BMI van de controlegroep, waaruit kan worden geconcludeerd dat de Gezonde Basisschool van de Toekomst gemiddeld leidt tot een lagere BMI. Als in plaats daarvan in een voor-navergelijking echter alleen naar de BMI van de interventiegroep over tijd zou zijn gekeken, zou er ten onrechte zijn geconcludeerd dat de Gezonde Basisschool van de Toekomst ineffectief is, omdat de BMI in de interventiegroep niet veranderde na de start van de interventie. Het betrekken van de controlegroep maakt echter duidelijk dat er iets anders aan de gang is: in dezelfde periode nam de BMI in de controlegroep namelijk juist toe. Zonder de Gezonde Basisschool van de Toekomst had de interventiegroep waarschijnlijk dezelfde stijging in BMI getoond als de controlegroep.

3.5 Voorbeelden

Aan de hand van voorbeelden laten we zien dat inzet van bewijsmethoden op ten minste SMS-niveau 3 op uiteenlopende preventiedomeinen haalbaar is. Tabel 3.1 geeft een overzicht van deze voorbeelden waarbij de adviescommissie per studie de methode heeft gescoord op de Maryland Scale. In Bijlage B geven we per voorbeeld een overzicht van de onderzoeksopzet, resultaten, bewijsmethode en externe validiteit. De onderzoeksopzet beschrijven we daarbij aan de hand van de Populatie, Interventie, Comparator, Outcome en Tijd (PICOT).

Tabel 3.1 Overzicht van voorbeelden

Nr	Naam	Schaal	Complexiteit	Methode	SMS-score	Studie
1	Samen gezond	Individu	Gestandaardiseerd	Difference-in-differences	3	Agachi et al. (2024)
2	Begeleiding bij stoppen met roken	Individu	Gestandaardiseerd	(Systematische review van) RCT's en natuurlijke experimenten	5	Lancaster & Stead (2017)
3	Conditional cash transfers in Brazilië	Individu	Gestandaardiseerd	Difference-in-differences	3	Cavalcanti et al. (2025)
4	Groepsinterventie tegen stress in de gezondheidssector	Individu	Gestandaardiseerd	Difference-in-differences	3	Dawson et al. (2021)
5	Griepvaccinatie	Individu	Gestandaardiseerd	(Systematische review van) RCT's en observationele studies	3+	Ferdinands et al. (2024)
6	Bevolkingsonderzoek longkanker	Individu	Gestandaardiseerd	RCT	5	de Koning t al. (2020)
7	Op Volle Kracht	Individu	Gestandaardiseerd	Cluster-RCT	5	de Jonge-Heesen et al. (2020)
8	Grip en Glans	Individu	Maatwerk	RCT	5	Kremers et al. (2006)
9	Preventie van onnodig zorggebruik	Individu	Maatwerk	Difference-in-differences	3	Sevak et al. (2018)
10	Myopie-preventie	Collectief	Gestandaardiseerd	Cluster-RCT	5	Wu et al. (2018)
11	Minimumleeftijd alcoholconsumptie	Collectief	Gestandaardiseerd	Regression discontinuity design + difference-in-differences	4	Carpenter & Dobkin (2011)
12	Interventie tegen bingedrinken onder scholieren	Collectief	Gestandaardiseerd	Cluster-RCT	5	Vargas-Martinez et al. (2019)
13	Verhoging suikertaks	Collectief	Gestandaardiseerd	Difference-in-discontinuities	4	Øvrebo et al. (2020)
14	Milieuzones	Collectief	Gestandaardiseerd	(Systematische review van) difference-in-differences studies	3	Margaryan (2021)
15	Gezonde Basisschool van de Toekomst	Collectief	Maatwerk	Difference-in-differences	3	Bartelink et al. (2019)
16	Jump-In-schoolprogramma	Collectief	Maatwerk	Difference-in-differences	3	Busch et al. (2023)
17	Stedelijke vernieuwing	Collectief	Maatwerk	Difference-in-differences	3	Ruijsbroek et al. (2017)
18	Workplace wellness programs	Collectief	Maatwerk	Cluster-RCT	5	Song & Baicker (2019)
19	Head start	Collectief	Maatwerk	Regression discontinuity design	4	Ludwig & Miller (2007)
20	WHO STOPS	Collectief	Maatwerk	Stepped wedge cluster-RCT	5	Allender et al. (2021)

Noot: Deze tabel scoort twintig studies op de Maryland Scale en vat ook een aantal andere kenmerken samen, een uitgebreide toelichting staat in Bijlage B. Het nummer in de eerste kolom volgt de nummering in deze bijlage. De kolom Schaal maakt onderscheid naar interventies die aan individuen worden gegeven of op collectief niveau worden toegediend, bijvoorbeeld scholen, wijken, bedrijven of de hele samenleving. Een interventie die zowel individuele als collectieve elementen bevat delen we ook in als collectief. De kolom Complexiteit maakt onderscheid naar of iedereen dezelfde behandeling krijgt of dat er sprake is van maatwerk.

De voorbeelden omvatten uiteenlopende domeinen. Onderzoeken naar griepvaccinatie (5) en bevolkingsonderzoek naar longkanker (6) vallen onder medische preventie. Onderzoek naar aanpakken zoals de begeleiding bij stoppen met roken (2) of het tegengaan van eenzaamheid en gebrek aan zelfregie onder vrouwen van boven de 50 (8) hebben betrekking op leefstijl. Onderzoek naar de effectiviteit van milieuzones (14) en stedelijke vernieuwing (17) hebben betrekking op leefomgeving. Andere onderzoeken gaan over interventies op de werkvloer (4, 18), mentale gezondheid (7) of jeugdpreventie (10). We hebben ook voorbeelden opgenomen van preventiebeleid door accijnzen (13) of regulering (11).

We kunnen binnen de voorbeelden ook een onderscheid maken naar individuele versus collectieve interventies en naar gestandaardiseerde interventies en maatwerk. Deze verschillen zijn relevant, omdat het voor collectieve en maatwerkinterventies lastiger kan zijn om effecten vast te stellen ten opzichte van een geloofwaardige controlegroep. Bij collectieve interventies op scholen, zoals de Gezonde Basisschool van de Toekomst (15), zullen andere scholen bijvoorbeeld als controlegroep moeten fungeren. Bij een maatregel op het niveau van de gehele populatie, zoals bij de minimumleeftijd voor alcoholconsumptie (11), is een ander design nodig om toch tot een controlegroep te komen.

Maatwerkinterventies onderscheiden zich van gestandaardiseerde interventies doordat ze uit een combinatie van verschillende elementen bestaan en doordat de toepassing hiervan afhankelijk is van de context. Hieronder vallen ook programma's die bestaan uit meerdere interventies en programma's die gestuurd worden vanuit een bepaald doel, bijvoorbeeld het realiseren van een whole of government approach, of een transitie naar een gezonde omgeving voor kinderen om op te groeien. Een voorbeeld van die laatste categorie is de Whole-of-Systems (WHO) aanpak voor kindero obesitas met een *community based*-interventie (20). Dit programma is ontwikkeld door relevante actoren in de betreffende gemeenschap en bestaat uit veel verschillende activiteiten, van regels aan ouders om alleen gezonde dranken mee te geven aan hun kinderen op school, tot de aanleg van fietspaden om fysieke activiteit te stimuleren.

Ondanks deze grote verschillen rapporteert Tabel 3.1 voor al deze preventiedomeinen en typen maatregelen onderzoek dat bewijsmethoden gebruikt op ten minste SMS-niveau 3. Veel studies scoren zelfs hoger. Ondanks de door de TWG-KBP gesignaleerde belemmeringen voor toepassing van een RCT voor preventie vinden we zelfs een groot aantal RCT's. Een kanttekening bij deze tabel is dat we hebben gescoord op de methode zonder te controleren of ook aan alle voorwaarden is voldaan. Niet alle studies met een *difference-in-differences*-aanpak laten bijvoorbeeld zien dat pretrends parallel lopen. Deze studies hebben we hier op niveau 3 gescoord, maar in Bijlage B benoemen we eventuele tekortkomingen bij de discussie van de methode.

4 Toepasbaarheid

De toepassing van onderzoeksresultaten uit een andere context vereist dat de onderzoeks- en implementatiecontext vergelijkbaar zijn op kenmerken die van invloed zijn op de werkzaamheid van een interventie. Een goede inschatting van de toepasbaarheid van onderzoek vereist daarom zicht op de werkzame elementen. Een minimale vereiste hiervoor is een theoretische onderbouwing van waarom en hoe de interventie tot de beoogde resultaten moet leiden.

De effectiviteit van een maatregel moet bij een ex-ante beoordeling vaak beoordeeld worden op basis van onderzoek uit een andere context. Het kan bijvoorbeeld gaan om een andere gemeente, een ander land of een experimentele context. Zoals in het vorige hoofdstuk beschreven, moeten die onderzoeken intern valide zijn, maar dat betekent nog niet dat ze toepasbaar zijn op de te beoordelen maatregel. De resultaten moeten ook toepasbaar zijn in de context waarin de maatregel wordt geïmplementeerd. Dit hoofdstuk gaat over de vraag wanneer aan deze voorwaarde is voldaan.

Centrale begrippen in dit hoofdstuk zijn context, implementatie en werkzame elementen. De context van het onderzoek gaat over de factoren die van invloed zijn op de effecten van de maatregel. De context van de te beoordelen maatregel, ofwel de implementatiecontext, moet hierbij aansluiten. Om te kunnen beoordelen of verschillen tussen deze twee contexten relevant zijn, moet het onderzoek inzicht geven in wat de maatregel effectief maakt. Wat zijn de werkzame elementen? Verschillen tussen de onderzoeks- en implementatiecontext zijn namelijk alleen relevant als ze gevolgen hebben voor de werkzaamheid van de maatregel.

Deze nadruk op de onderzoeks- en implementatiecontext en de werkzame elementen sluit goed aan bij richtlijnen voor de evaluatie van complexe interventies (Thomas et al., 2019; Pfadenhauer et al., 2017). Ook bij medische interventies wordt toepasbaarheid meegewogen bij de beoordeling van bewijs. De GRADE-methodiek kijkt bijvoorbeeld naar hoe direct de PICOTs van het onderzoek en de te beoordelen interventie op elkaar aansluiten (Schünemann et al., 2025). Aspecten van implementatie komen hierbij aan de orde in de vertaalslag van bewijs naar aanbevelingen. Het Cochrane-handboek licht toe hoe complexe interventies binnen deze methodiek kunnen worden beoordeeld en verwijst daarbij naar dezelfde richtlijnen voor de evaluatie van complexe interventies (Thomas et al., 2019).

In de volgende paragraaf gaan we nader in op context en werkzame mechanismen. We lichten de beoordeling van toepasbaarheid toe aan de hand van voorbeelden. Hierbij gaan we uit van het perspectief van de beoordelaar die de onderzoeks- en implementatiecontexten vergelijkt bij de beoordeling van het bewijs. Dit verschilt van de bespreking van passende bewijsmethoden in Hoofdstuk 3 waarbij de focus ligt op de beoordeling van individuele onderzoeken. Paragraaf 4.2 bespreekt hoe individuele onderzoeken deze beoordeling kunnen faciliteren. De samenhang tussen bewijsmethode en toepasbaarheid lichten we nader toe in Box 4.1.

Box 4.1: Van bewijsmethode naar toepasbaarheid

Interne validiteit is een noodzakelijke voorwaarde voor toepasbaarheid. Als de effectiviteit in de onderzoekscontext al niet overtuigend is aangetoond, mogen we er ook niet van uitgaan dat de resultaten geldig zijn in de implementatiecontext. Bewijsmethoden met een lagere interne validiteit kunnen wel een grotere toepasbaarheid hebben, bijvoorbeeld omdat ze gebaseerd zijn op observationele data in plaats van op een gecontroleerd experiment. De echte wereld (real world) waarin deze data verzameld zijn kan namelijk dichterbij de implementatiecontext staan dan een experimentele context. Dit kan een reden zijn om bij de effectbeoordeling uit te gaan van observationeel onderzoek, zolang de gebruikte methode scoort op ten minste niveau 3 van de Maryland Scale. Daar staat tegenover dat sommige observationele methoden effecten voor een specifieke groep binnen de onderzoekspopulatie meten, zoals bij Instrumentele Variabelen (zie Paragraaf 3.2). Dat beperkt juist de toepasbaarheid.

Als de beoordelaar beschikt over meerdere intern valide onderzoeken die een effect in verschillende contexten aantonen, maakt dit de toepasbaarheid van de resultaten in de implementatiecontext aannemelijker. De beschikbaarheid van meerdere studies compenseert echter niet voor tekortschietende interne validiteit. Ook als er meerdere studies bijvoorbeeld een effect vinden zonder daarbij een controlegroep te gebruiken is er dus nog geen sprake van passend bewijs. Die studies kunnen namelijk allemaal om dezelfde reden vertekend zijn, bijvoorbeeld omdat de meest gemotiveerde mensen zich aanmelden voor een interventie.

4.1 Context en werkzame elementen

De context van een interventie bestaat uit de factoren die de interventie beïnvloeden en gevolgen hebben voor de effectiviteit. Context omvat verschillende aspecten, bijvoorbeeld geografisch, epidemiologisch, sociocultureel, sociaaleconomisch, ethisch, juridisch en politiek.³⁰ Kenmerken van de populatie maken hier deel van uit. Als factoren die invloed hebben op de effectiviteit van een interventie verschillen tussen het onderzoek en de implementatie, dan kan dit de toepasbaarheid van de onderzoeksresultaten in een bepaalde context beperken. Een studie die laat zien dat het dragen van fietshelmen leidt tot een daling van letsel als gevolg van fietsongelukken zou bijvoorbeeld weleens beperkt toepasbaar in Nederland kunnen zijn, als deze is uitgevoerd in een land waar nauwelijks gescheiden fietspaden zijn. Wij hebben namelijk veel gescheiden fietspaden waardoor de kans op ernstig letsel sowieso lager is.

Om de verschillen in context goed te kunnen meewegen bij het beoordelen van de toepasbaarheid van de resultaten van een studie in een andere context, moeten deze in samenhang met het mechanisme worden bekeken dat de werkzaamheid van de preventieve maatregel verklaart. Men spreekt ook wel van werkzame elementen. Dit zijn de elementen die er al dan niet in onderlinge samenhang voor zorgen dat de interventie tot het effect leidt. Werkzame elementen van de gecombineerde leefstijlinterventie (GLI) zijn bijvoorbeeld de lange duur van de interventie, de multidisciplinaire aanpak en de toepassing van gedragsveranderingstechnieken zoals motiverende gespreksvoering.

We illustreren de samenhang tussen context en werkzame elementen aan de hand van een maatregel voor preventie van vapes bij jongeren waarvan de effecten getest zijn op een populatie van 12 tot 16 jaar. Zijn de resultaten hiervan toepasbaar op volwassenen die vapes? Een analyse van

³⁰ Deze indeling is gebaseerd op de CICI-richtlijn (Pfadenhauer et al., 2017). Daarnaast onderscheidt deze richtlijn nog het begrip setting dat verwijst naar de fysieke locatie waar de interventie plaatsvindt. Zie voor een overzicht van de literatuur over toepasbaarheid bijvoorbeeld Burchett et al. (2018) en Schloemer en Schröder-Bäck (2018).

de werkzame elementen kan deze vraag beantwoorden. Als de benadering door leerkrachten een van de werkzame elementen is, dan sluit dat bijvoorbeeld niet aan bij de oudere doelgroep. Maar de onderzochte interventie kan mogelijk wel toegepast worden voor jongeren van 12 tot 16 jaar ter preventie van roken van sigaretten. Een voorwaarde hiervoor is dat de mechanismen onder beginnen met vaperen en beginnen met roken vergelijkbaar zijn en ook bijvoorbeeld de benadering door leerkrachten in beide gevallen hetzelfde is.

Die werkzame mechanismen en de invloeden vanuit de context kunnen enerzijds sociaal van aard zijn, zoals in het voorbeeld hierboven van de rol van leerkrachten, of fysiek, zoals in het voorbeeld van fietshelmen. Maar er zijn ook veel andere invloeden denkbaar:

- **Cultuur.** Bijvoorbeeld bij onderzoek naar de effectiviteit van een voedingsinterventie. Een studie over de effectiviteit van een interventie bij mensen met een cultuur waarin eten sterk gekoppeld is aan samenzijn, bijvoorbeeld door hen bewust te maken van de sociale druk om te eten, is waarschijnlijk niet goed toepasbaar op bevolkingsgroepen met een meer individualistische cultuur. Of, om een ander voorbeeld te geven: Onderzoek naar de deelname aan vaccinatieprogramma's in een streng religieuze populatie is vermoedelijk niet goed toepasbaar in een meer seculiere omgeving.
- **Gezondheidsvaardigheden.** Dit betreft het vermogen om kennis over gezondheid te verkrijgen, te begrijpen en toe te passen. Dit speelt bij preventieve maatregelen vaak een rol, vooral wanneer het over gezondheidsgedrag gaat. Een studie naar een educatieve maatregel, bijvoorbeeld rondom voeding, die is getest onder studenten die onderwijs op academisch niveau volgen, is daarom waarschijnlijk niet goed toepasbaar op een populatie met beperkte gezondheidsvaardigheden.
- **Epidemiologie en ander preventiebeleid.** Hoe vaak een ziekte of aandoening voorkomt maakt ook uit voor de effecten van een interventie. Tienerzwangerschappen komen in de Verenigde Staten bijvoorbeeld veel vaker voor dan in Nederland. Amerikaans onderzoek naar de effecten van een interventie om tienerzwangerschappen te beperken door gratis anticonceptiemiddelen uit te delen kan daarom niet zonder meer toegepast worden in Nederland. Verschillen in preventie kunnen hierbij ook een rol spelen. Wanneer die anticonceptiemiddelen in Nederland sowieso al meer worden gebruikt, of wanneer de preventieve zorg die in Nederland gegeven wordt om zwangerschappen te voorkomen al veel beter is, beperkt dit ook de toepasbaarheid. In dat geval is het staande preventiebeleid mogelijk een substituuut, maar preventiebeleid kan ook complementair zijn. Zo zijn interventies om overgewicht onder jongeren te beperken mogelijk effectiever in een context waarin ongezond eten ook door regulering en belastingen wordt ontmoedigd. Resultaten ervan zijn dan minder toepasbaar in een context waarin dat niet het geval is.
- **Zorgstelsel.** Ook aspecten van de organisatie van zorg, zoals het wel of niet hanteren van eigen bijdragen of de plaats waar zorg wordt aangeboden, kunnen de toepasbaarheid van interventies die in een ander zorgsysteem zijn getest beïnvloeden. Zo zal de deelnamebereidheid van ouders voor een vaccinatieprogramma voor hun kinderen afhangen van de manier waarop deze wordt aangeboden, bijvoorbeeld door de huisarts, via het consultatiebureau of in massavaccinatiecampagnes. Buitenlandse studies waarin vaccinaties gegeven worden door kinderartsen zijn daarmee niet direct vertaalbaar naar de Nederlandse context.

Behalve van de context waarin de maatregel zal worden toegepast, in samenhang met het werkzame mechanisme onderliggend aan die maatregel, is het ook belangrijk een beoordeling te maken van de condities die nodig zijn voor een goede implementatie. Denk aan de beschikbaarheid van

competente arbeidskrachten om de interventie uit te voeren, of voldoende financiële middelen. Wanneer de condities waaronder de maatregel in een onderzoek succesvol is gebleken niet beschikbaar zijn in de context waarin deze wordt toegepast, dan kan dat ertoe leiden dat de geschatte effectiviteit in die context niet wordt gehaald. Als een interventie ter verhoging van therapietrouw in apotheken bijvoorbeeld is uitgevoerd door apothekersassistenten, en die assistenten in het land waarin ze zijn onderzocht een veel meer gespecialiseerde opleiding hebben genoten dan in Nederland, dan kan dat ertoe leiden dat de interventie in de Nederlandse context veel minder effectief is. Door de apothekersassistenten een training te geven die aansluit bij het opleidingsniveau in Nederland kan de effectiviteit mogelijk wel weer verhoogd worden.

De bespreking van de twintig voorbeeldonderzoeken uit Tabel 3.1 in Bijlage B gaat ook in op toepasbaarheid. Zo draagt het gebruik van observationele data bij aan de toepasbaarheid van een aantal onderzoeken. Bij andere onderzoeken worden concrete randvoorwaarden voor toepasbaarheid benoemd. Zo is de toepasbaarheid van een onderzoek naar longkankerscreening afhankelijk van de ontwikkeling van het zorgsysteem, de deelname, rokersprofielen, bevolkingssamenstelling en kosten en acceptatie van screening. Verder gaat Bijlage B bij elk voorbeeldonderzoek ook in op de beschikbaarheid van andere studies, wat de toepasbaarheid vergroot.

4.2 Wat vraagt dit van het onderzoek?

Er zijn meerdere manieren om zicht te krijgen op de context en werkzame elementen. Een minimale vereiste is dat de interventie theoretisch onderbouwd is. Er moet een interventielogica of een programma- of beleidstheorie zijn die uitlegt waarom en hoe de interventie tot de beoogde uitkomsten leidt. De werkzame elementen, de mechanismen waardoor deze elementen tot uitkomsten leiden en de rol van de context maken hier deel van uit.

Richtlijnen voor de evaluatie van complexe interventies adviseren om de theorie al bij de beginfase van de ontwikkeling van een interventie op te stellen en vervolgens te verfijnen naarmate het begrip van de werking van de interventie en de interactie met de context toenemen (Pfadenhauer et al., 2017; Skivington et al., 2021). Een van de methoden hiervoor is ‘realist evaluation’ waarbij hypothesen over de relatie tussen context, mechanismen en uitkomsten gedurende het onderzoek iteratief worden getoetst en verfijnd (Pawson et al., 2005).

Een procesevaluatie die kwantitatieve gegevens met kwalitatieve informatie combineert kan ook zicht geven op de werkzame elementen en mechanismen. Dit is van belang bij complexe interventies, maar kan ook zinvol zijn voor gestandaardiseerde interventies. Vragen in een procesevaluatie zijn hoe de interventie precies vorm krijgt en of het ontwerp hierbij goed wordt toegepast, hoe de interventie tot veranderingen leidt en welke rol context hierin speelt. De MRC heeft een speciale richtlijn opgesteld voor procesevaluaties van complexe interventies (Moore et al., 2015).

Richtlijnen voor de evaluatie van complexe interventies geven ook aan hoe implementatie hierin kan worden meegenomen. De MRC-richtlijn beveelt aan om implementatie vanaf het begin in alle fases van het onderzoek mee te nemen. Zo moet de evaluatie inzicht geven in de invloed van contextfactoren op de adoptie en het bereik van de interventie en moet de interventielogica of programma- of beleidstheorie inzicht geven in welke werkzame elementen behouden moeten blijven en waar aanpassing mogelijk is. Verder benadrukt de richtlijn het belang van het betrekken van stakeholders en doelgroep gedurende het hele proces.

Ook kwantitatief onderzoek kan bijdragen aan begrip van de mechanismen die leiden tot de uitkomsten. De hiervoor gebruikte methode staat bekend als ‘mediation analysis’ waarbij

uitkomstvariabelen in de statistische analyse worden vervangen door intermediaire uitkomsten. Box 4.2 beschrijft hoe deze methode heeft bijgedragen aan een beter begrip van de werkzaamheid van de interventie Grip en Glans.

Box 4.2 Werkzame elementen van Grip en Glans in beeld

Grip en Glans is een interventie onder vrouwen van boven de vijftig met een verminderd subjectief welbevinden. In een aantal wekelijkse bijeenkomsten wordt in groepsverband gewerkt aan het ontwikkelen van zelfmanagementvaardigheden. Het veronderstelde mechanisme is dat verbetering van zelfmanagementvaardigheid leidt tot een hoger welbevinden. Dit is gebaseerd op de theorie van Zelfmanagement van welbevinden.³¹

Goedendorp en Steverink (2017) toetsen het verwachte mechanisme (mediatie-effect) in twee stappen. Stap 1 kijkt of de veranderingen in zelfmanagementvaardigheid en welbevinden significant groter zijn in de interventiegroep dan in de controlegroep. Stap 2 kijkt of de veranderingen in welbevinden toegeschreven kunnen worden aan veranderingen in zelfmanagementvaardigheid.

Het onderzoek vindt een volledige en significante mediatie, wat inhoudt dat de verbetering in welbevinden verklaard wordt door de verbetering in zelfmanagementvaardigheid die door de interventies teweeg wordt gebracht. Dit is een ondersteuning voor de theorie van Zelfmanagement van welbevinden die ten grondslag ligt aan alle Grip en Glans-interventies (Steverink, 2009; 2014) en stelt dat het vergroten van zelfmanagementvaardigheid resulteert in een verbeterd welbevinden.

Soms is het mogelijk om de gevolgen van een slechte aansluiting van effectiviteitsstudies op de implementatiecontext kwantitatief in beeld te brengen en te beperken, namelijk door het effect van een maatregel te schatten voor deelgroepen (Schünemann et al., 2025). In het voorbeeld uit de vorige paragraaf over hoe de effectiviteit van een voedingsinterventie samen kan hangen met cultuur kan het effect geschat worden voor deelgroepen met verschillende herkomst. Dit geeft een beeld van de gevoeligheid voor herkomst, en de effectschatting voor de groep die het meest lijkt op de doelgroep voor de te beoordelen maatregel kan als uitgangspunt dienen. Hiervoor moet het onderzoek wel voldoende waarnemingen binnen de relevante deelgroepen hebben.

Het merendeel van de twintig voorbeeldonderzoeken uit Tabel 3.1 bevat een beleidstheorie. Voor sommige interventies is ook een procesevaluatie beschikbaar, zoals voor de Gezonde Basisschool van de Toekomst. Bijlage B licht dit nader toe.

³¹ Zie [het werkblad](#) van de erkende interventie.

5 Langetermijneffecten

Een belangrijk deel van de baten van preventie ontstaat pas op de lange termijn, maar dit hoeft geen belemmering te zijn voor passend bewijs. Intermediaire uitkomsten of kortetermijneffecten van preventie kunnen worden doorvertaald naar de lange termijn, mits de gehanteerde methode voor de doorvertaling empirisch gevalideerd is en de aannames en de gevoeligheid hiervoor expliciet worden gemaakt.

Een belangrijk deel van de baten van preventie ontstaat pas op de lange termijn. Zo verschijnen de gezondheidseffecten van een suikertaks, zoals minder diabetes en hart- en vaatziekten, vaak pas vele jaren na invoering. Om tot een betrouwbare effectbepaling te komen zouden we deze langetermijneffecten het liefst op eenzelfde manier meten als de kortetermijneffecten, in een meerjarige monitor en met behulp van de in Hoofdstuk 3 besproken bewijsmethoden. In de praktijk is dit echter zelden haalbaar: een gerandomiseerd experiment zou tientallen jaren moeten lopen om uitspraken te kunnen doen over de langetermijneffecten. De meeste onderzoeken brengen daarom de effecten na hooguit enkele jaren in kaart.

Om toch uitspraken te kunnen doen over de langetermijneffecten van preventie kunnen deze kortetermijneffecten vertaald worden naar langetermijneffecten. Hierbij kan gebruikgemaakt worden van intermediaire uitkomsten. Dit zijn andere uitkomsten dan de (beoogde) einduitkomsten die worden gebruikt als vervangende indicatoren voor de (beoogde) einduitkomsten en waarvan het aannemelijk is dat deze doorwerken op de (beoogde) einduitkomst. BMI kan bijvoorbeeld dienen als intermediaire uitkomst voor de einduitkomst kans op diabetes type 2. De doorvertaling van intermediaire uitkomsten of kortetermijneffecten naar langetermijneffecten gebeurt via modellering. Dit kan zowel heel eenvoudig als heel complex worden vormgegeven, Box 5.1 licht dit toe.

Box 5.1 Wat is een model?

Met een model bedoelen we in deze richtlijn iedere impliciete of expliciete aanpak om gezondheidseffecten te extrapoleren naar een langere periode dan de studieduur. Het betreft dus niet noodzakelijkerwijs complexe mathematische systemen.

Voorbeelden

- Een intensief leefstijlprogramma voor mensen met overgewicht en diabetes leidt bijvoorbeeld tot een gewichtsverlies van 4,7 procent na 4 jaar (Look AHEAD Research Group, 2010). Vervolgens wordt de kosteneffectiviteit van het programma in kaart gebracht en moeten deze effecten worden doorvertaald naar de lange termijn. De onderzoekers hanteren daarbij een zeer eenvoudig model: ze gaan ervanuit dat het gemiddelde gewichtsverlies van 4,7 procent aanhoudt over de rest van het leven.
- In veel klinische trials worden effecten van een interventie op een bepaalde uitkomst over de tijd (bijvoorbeeld sterfte) geschat met duurmodellen. De geobserveerde tijdsperiode is vaak relatief kort (bijvoorbeeld 1 jaar). Effecten worden daarna op basis van het geschatte model over de rest van het leven geëxtrapoleerd.
- Om de relatie tussen geobserveerde intermediaire uitkomsten (bijvoorbeeld BMI) en gezondheidsuitkomsten op de lange termijn (bijvoorbeeld prevalentie van diabetes) te berekenen wordt vaak gebruikgemaakt van Markov-transitiemodellen. Hierbij is BMI een relatieve risicofactor voor het krijgen van diabetes op een bepaald moment in het leven. Deze wordt gekwantificeerd op basis van de bestaande literatuur (zie bijvoorbeeld Van der Ende et al., 2024).

De doorvertaling naar de lange termijn brengt onzekerheid met zich mee. Verschillende andere richtlijnen suggereren daarom om bewijs dat gebruikmaakt van modellering lager te scoren dan op basis van de gehanteerde methode voor het in kaart brengen van de tussenuitkomst of kortetermijnuitkomst gedaan zou worden.³² Deze richtlijn erkent echter dat modellering soms onvermijdelijk is en stelt dat er dan nog steeds sprake kan zijn van passend bewijs. Het uitgangspunt is wel dat het model empirisch gevalideerd is en dat aannames en de gevoeligheid van de uitkomsten hiervoor expliciet worden gemaakt. Daarnaast geldt ook voor effecten op intermediaire uitkomsten of kortetermijnuitkomsten dat deze ten minste op niveau 3 van de Maryland Scale aangetoond moeten zijn (Hoofdstuk 3) en dat ze toepasbaar zijn in de implementatiecontext (Hoofdstuk 4).

Modellering van de langetermijneffecten is daarmee nog geen volwaardig substituuut voor een effectmeting. GRADE adviseert om alleen gebruik te maken van intermediaire uitkomsten wanneer er geen bewijs beschikbaar is over de einduitkomsten (Schünemann et al., 2013). Waar mogelijk heeft een effectmeting op de lange termijn dan ook de voorkeur.³³ Dit verhoogt de kwaliteit van het bewijsmateriaal ten opzichte van modellering. Wanneer er dan toch gebruik wordt gemaakt van een model, dan is het aan te bevelen om ook na modellering de effecten van een interventie regelmatig te blijven monitoren en evalueren.

In het vervolg van dit hoofdstuk gaan we allereerst nader in op de onzekerheden die ontstaan bij modellering. Vervolgens gaan we in op de eisen die de richtlijn stelt ten aanzien van intermediaire uitkomsten, empirische validatie en het in kaart brengen van onzekerheid. We doen ook handreikingen voor hoe dit vormgegeven kan worden.

5.1 Bronnen van onzekerheid

De doorvertaling van korte naar lange termijn brengt onzekerheid met zich mee. Om op basis van kortetermijnresultaten uitspraken te doen over de langetermijneffecten moeten namelijk aannames gemaakt worden over hoe het gevonden effect zich in de tijd ontwikkelt. Er zijn verschillende oorzaken waardoor het kortetermijneffect kan verschillen van het langetermijneffect:

- **Het beoogde effect treedt lang na blootstelling aan de interventie op**
De gezondheidseffecten van een suikertaks doen zich bijvoorbeeld vaak pas vele jaren na de invoering van de suikertaks voor. Op de korte en middellange termijn kunnen wel effecten op prijs, consumptie van suikerhoudende producten, suikerconsumptie en gewicht worden gemeten (zie bijvoorbeeld Chen et al., 2009; Øvrebo et al., 2020).
- **Het effect van de interventie dooft uit**
Een interventie dooft uit als de effecten die in eerste instantie door de interventie zijn bereikt in de loop van de tijd verminderen of helemaal verdwijnen, zodra de interventie stopt of minder intensief wordt toegepast. Een intensief leefstijlprogramma voor mensen met overgewicht en diabetes leidde bijvoorbeeld tot een gewichtsverlies van 8,6 procent na één

³² Het gebruik van intermediaire uitkomsten of kortetermijnuitkomsten leidt bijvoorbeeld tot verlaging van de bewijskwaliteit in GRADE (McCaul et al., 2025). De WHO stelt zelfs voor om resultaten op basis van modellen op het laagste bewijsniveau van GRADE in te schalen.

³³ Het gebruik van observationele onderzoeksmethoden kan daarbij soms uitkomst bieden, omdat het mogelijk is om (ver) terug te kijken in de tijd. Gracner et al. (2024) maken bijvoorbeeld gebruik van een natuurlijk experiment met suikerrantsoenering uit 1953 om de langetermijneffecten van suikerconsumptie van (zwangere) vrouwen en jonge kinderen te analyseren. Ze volgen daarvoor de kinderen tot hun 65^{ste} levensjaar. Hieruit concluderen ze dat kinderen waarvan de moeders minder suiker consumeerden of die zelf in hun eerste levensjaren minder suiker consumeerden op latere leeftijd minder vaak diabetes type 2 en hypertensie ontwikkelden.

jaar, maar na vier jaar bleek het gewichtsverlies nog maar 4,7 procent (Look AHEAD Research Group, 2010).

- **Door andere (gedrags)effecten wordt het effect sterker of zwakker**

Gedragsreacties of andere effecten kunnen interfereren met het effect van de interventie. Deze effecten kunnen in principe zowel op de korte als lange termijn plaatsvinden, maar het risico wordt groter naarmate de tijd tussen de relevante uitkomstmaat en de interventie toeneemt. Zo leidt een daling van het aantal mensen met overgewicht in eerste instantie tot minder zorgkosten omdat de kans op aan overgewicht gerelateerde ziekten daalt. Op de langere termijn stijgen de zorgkosten echter doordat mensen langer leven en daarmee de kans op het krijgen van andere ziekten toeneemt (Van Baal et al., 2008). Mensen die als gevolg van een suikertaks gewicht verliezen zouden op termijn ander leefstijlgedrag kunnen verslechteren (bijvoorbeeld meer roken). Of fabrikanten van vruchtensappen reageren op invoering van de suikertaks door een beetje zuivel toe te voegen aan hun receptuur. Ook dit beïnvloedt de langetermijneffecten van de suikertaks.

Wanneer onderzoek niet direct de (beoogde) einduitkomst meet, maar wel effecten op intermediaire uitkomsten, brengt de doorvertaling naar de lange termijn additionele onzekerheid met zich mee. Als er gebruik wordt gemaakt van intermediaire uitkomsten moeten er namelijk niet alleen aannames gedaan worden over de ontwikkeling van het effect over de tijd, maar ook over de relatie tussen de intermediaire uitkomst en de einduitkomst.³⁴

5.2 Intermediaire uitkomsten

Niet elke mogelijke tussenuitkomst leidt automatisch tot passend bewijs. Deze richtlijn vereist dat het effect op de einduitkomst met een bewijsmethode van ten minste niveau 3 op de Maryland Scale is vastgesteld en dat er een theoretische onderbouwing van deze relatie bestaat.³⁵ Daarmee hanteert de richtlijn hetzelfde uitgangspunt voor de relatie tussen de tussen- en einduitkomst als voor de relatie tussen de interventie en de tussenuitkomst. Deze minimumeis wordt in de literatuur al vaak gehaald (zie bijvoorbeeld Ciani et al., 2021).

Bij de twintig voorbeeldstudies in Bijlage B komen verschillende voorbeelden van intermediaire uitkomsten aan bod. In het onderzoek van Lancaster en Stead (2017) wordt bijvoorbeeld het effect geschat van stoppen-met-rokenbegeleiding op de kans om te stoppen met roken. Dit is niet het uiteindelijke beoogde effect van de interventie, dat is een betere gezondheid. Maar dat effect treedt pas op de lange termijn op. De relatie tussen roken en gezondheid is echter in andere literatuur veelvuldig aangetoond, wat dit een valide intermediaire uitkomst maakt. Vargas-Martínez et al. (2019) onderzoeken het effect van een interventie tegen bingedrinken onder middelbare scholieren op het aantal bingedrinksessies van deze jongeren. Ook hier is de beoogde einduitkomst een betere gezondheid, wat pas op de langere termijn optreedt. De relatie tussen bingedrinken en gezondheid is echter in eerder onderzoek in kaart gebracht. Er bestaan daarnaast ook standaardlijsten van veelgebruikte intermediaire uitkomsten.³⁶

De richtlijn beveelt bovendien aan om te werken met intermediaire uitkomsten die zo dicht mogelijk bij het uiteindelijke gezondheidseffect liggen. Dit beperkt namelijk de onzekerheid rondom de extrapolatie. Vaak heeft een interventie via verschillende opeenvolgende intermediaire uitkomsten doorwerking op de langetermijngezondheid (zie Figuur 5.1). Zo werkt een suikertaks via meerdere

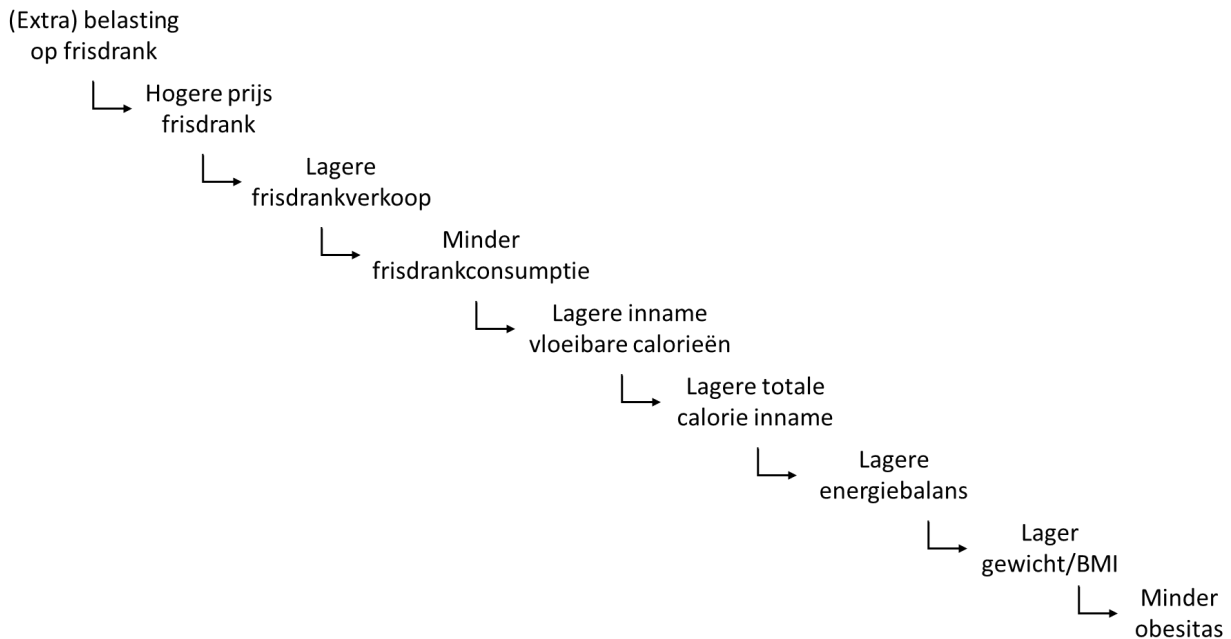
³⁴ In GRADE wordt daarom geadviseerd om bewijs op basis van intermediaire uitkomsten lager te beoordelen dan bewijs op basis van einduitkomsten gemeten op de korte termijn (Guyatt et al., 2011).

³⁵ Voor theoretische literatuur over de criteria voor intermediaire uitkomsten, zie o.a. Buyse en Molenberghs (1998), Prentice (1989) en Athey et al. (2025)

³⁶ De FDA houdt bijvoorbeeld een lijst bij van intermediaire uitkomsten op basis waarvan geneesmiddelen zijn goedgekeurd (FDA, 2025).

intermediaire uitkomsten door op de gezondheid: hogere prijzen en een lagere suikergehalte van drankjes leiden tot minder suikerconsumptie, een lagere calorie-inname, een lagere BMI en uiteindelijk een lager risico op diabetes type 2. Een studie die alleen het effect van de suikertaks op suikergehalte en -verkoop meet, moet aannemen in hoeverre de lagere consumptie wordt gecompenseerd door de inname van andere producten, of dat een lagere calorie-inname mogelijk wordt tenietgedaan door veranderingen in ander gedrag, zoals meer roken (Dickson et al., 2025). Studies die direct BMI meten, hoeven alleen aannames te maken over de relatie tussen BMI en de kans op diabetes type 2 (Liu et al., 2025).

Figuur 5.1 Relatie tussen suikertaks en obesitas via een keten van intermediaire uitkomsten



Bron: Briggs et al. (2013), bewerking adviescommissie

Een ander voorbeeld betreft een interventie gericht op het bevorderen van handhygiëne onder ziekenhuispersoneel ter preventie van infecties. In dit onderzoek wordt een intermediaire uitkomst gebruikt die zo ver van de beoogde einduitkomst af staat dat het niet aannemelijk is dat het gevonden effect betrouwbaar kan worden doorvertaald naar de einduitkomst (Moreno et al., 2025). Het onderzoek kijkt namelijk uitsluitend naar het gebruik van handalcohol. Hoewel het plausibel is dat meer gebruik samenhangt met minder ziekenhuisinfecties zijn er ook andere mechanismen denkbaar waardoor een hoger gebruik niet leidt tot minder infecties. Het gebruik kan bijvoorbeeld ook verhoogd worden door de handalcohol door de gootsteen te spoelen.

5.3 Empirische validatie

De keuze voor de wijze van modellering en de hierbij gemaakte aannames zijn bepalend voor de uitkomsten van het model. Een volledige omschrijving van eisen over modelkwaliteit valt buiten het bereik van deze richtlijn. We stellen wel de minimumeis dat het gebruikte model empirisch gevalideerd is. Dit sluit aan bij een aantal andere werkwijzers. Zo eist de MKBA-werkwijzer voor het sociaal domein (Koopmans et al., 2016) empirische validatie voor een model dat gebruikt wordt voor effectbepaling. De richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg van ZIN bespreekt verschillende eisen voor de validatie van modellen (Zorginstituut Nederland, 2024c). De International Society for Pharmacoeconomics (ISPOR) heeft ook een richtlijn geschreven voor het uitvoeren van modelstudies die ingaat op eisen ten aanzien van de validatie (Caro et al., 2014).

Er bestaat geen blauwdruk voor de empirische validatie van modellen om gezondheidseffecten te extrapoleren. We kunnen in deze richtlijn daarom geen eenvoudige regels opstellen over hoe die validatie eruit moet zien. In de wetenschappelijke literatuur worden echter wel verschillende validatiestappen besproken, onder meer door Kopec et al. (2010) en Vemer et al. (2016). Daarbij maakt de literatuur onderscheid naar de empirische validatie van:

- **Interne consistentie.** Genereert het model uitkomsten die in overeenstemming zijn met de theoretische mechanismen? Met andere woorden, genereert het model daadwerkelijk de relaties die de modelleers veronderstellen?
- **Interne validiteit.** Genereert het model outputs die consistent zijn met de inputs? Bijvoorbeeld: levert een model dat incidenties van de huidige Nederlandse bevolking gebruikt ook prevalenties op die vergelijkbaar zijn met die van de Nederlandse bevolking?
- **Vergelijking met andere modellen.** Zijn de resultaten van het model vergelijkbaar met resultaten van andere modellen?
- **Prestaties op andere data.** Kan het model ook patronen ontdekken in data die niet zijn gebruikt bij de ontwikkeling ervan, bijvoorbeeld in historische gegevens?
- **Voorspellende prestaties.** Kan het model de gevolgen van een nieuwe interventie betrouwbaar voorspellen? Vaak is dit lastig, omdat er doorgaans een grote tijdsvertraging zit tussen de invoering van beleid en het optreden van effecten. In sommige gevallen is het echter wel mogelijk, bijvoorbeeld wanneer het gaat om intermediaire of kortetermijnuitkomsten.

Een interessant voorbeeld van hoe structurele modelvalidatie georganiseerd kan worden, is de jaarlijkse Mount Hood Diabetes Challenge. Hierin toetsen groepen gezondheidseconomische diabetesmodelleers gezamenlijk hun modellen. Tijdens deze bijeenkomsten krijgen de deelnemende groepen eenzelfde opdracht die zij vervolgens ieder met hun eigen model oplossen. Zo kan bijvoorbeeld worden gekeken in hoeverre de modellen in staat zijn het ziekteverloop te voorspellen op basis van nieuwe data die niet gebruikt zijn voor de oorspronkelijke modelschatting (Palmer, 2013). Een ander type opdracht bestaat uit het doorrekenen van eenzelfde beleidsvraag, zoals de kosteneffectiviteit van een specifieke interventie (Altunkaya et al., 2024). Door de uitkomsten van verschillende modellen naast elkaar te leggen ontstaat inzicht in de zogenoemde structurele onzekerheid. De Mount Hood Diabetes Challenge is daarmee een structureel moment om veelgebruikte diabetesmodellen empirisch te valideren.

5.4 Onzekerheid

Een laatste uitgangspunt van deze richtlijn met betrekking tot modellering is dat de gemaakte aannames en de gevoeligheid hiervoor expliciet in kaart worden gebracht. Dit betekent dat met gevoeligheidsanalyses wordt getoond in welke mate de resultaten veranderen wanneer de onderliggende aannames iets worden aangepast. Specifieke aandachtspunten zijn daarbij:

- **Persistentie van het effect op intermediaire uitkomsten.** Wanneer bijvoorbeeld de langetermijneffecten van een leefstijlinterventie worden geanalyseerd, is het essentieel om te laten zien hoe gevoelig de uiteindelijke resultaten zijn voor de aanname over de duurzaamheid van gedragseffecten. Gedragsveranderingen kunnen immers in de loop van de tijd afnemen waardoor het uiteindelijke effect op gezondheidsuitkomsten kleiner kan uitvallen dan aanvankelijk gemeten.
- **Bewijsbasis voor de relatie tussen intermediaire uitkomsten en einduitkomsten.** Vaak wordt deze relatie gemodelleerd op basis van schattingen uit de literatuur, zoals het relatieve risico van een verhoogde BMI op het ontstaan van diabetes. Omdat dergelijke schattingen altijd onzekerheid met zich meebrengen, moet een studie duidelijk maken in welke mate de resultaten gevoelig zijn voor deze onzekerheid in de onderliggende parameters.

- **Aannames over gedragsreacties.** Bij het doorrekenen van de effecten van een suikertaks op de volksgezondheid is het bijvoorbeeld bepalend hoe sterk consumenten hun gedrag aanpassen aan prijsstijgingen. De prijselasticiteit is echter omgeven met onzekerheid. Daarom is het noodzakelijk dat modelstudies inzichtelijk maken hoe sterk de einduitkomsten afhangen van deze gedragsaannames.

Een voorbeeld van een onderzoek waarin de gevoeligheid voor gemaakte aannames duidelijk in kaart wordt gebracht, is het onderzoek van het RIVM naar de impact van een getrapte verbruiksbelasting op de verkoop van suiker via alcoholvrije dranken (limonade, frisdrank, vruchtensappen, alcoholvrij bier) (RIVM, 2023). In het onderzoek worden de effecten gemodelleerd van een hogere belasting op de verkoopvolumes van alcoholvrije dranken. Hiervoor worden verschillende aannames gemaakt, waaronder over de relatie tussen de prijs van en vraag naar alcoholvrije dranken (prijselasticiteit), de mate waarin de hogere belasting wordt doorvertaald naar consumentenprijzen (*pass-through rate*) en de veranderingen in samenstelling van alcoholvrije dranken als gevolg van de belasting (herformulering naar dranken met een lager suikergehalte). De waarden voor deze variabelen worden gebaseerd op bevindingen uit de literatuur en de onderzoekers laten in verschillende gevoeligheidsanalyses zien wat de impact op de geschatte verkoop van suikerhoudende dranken zou zijn als er voor de verschillende variabelen net andere waarden waren genomen. Hiermee wordt expliciet gemaakt in welke mate de modeluitkomsten afhankelijk zijn van de gehanteerde aannames.

6 Aanbevelingen

Deze richtlijn stelt haalbare minimumeisen aan het bewijs dat nodig is om preventiebeleid passend te onderbouwen. Dit mag experimenten met nieuwe benaderingen niet in de weg staan.

Onderzoeksbekostiging en data-infrastructuur zijn randvoorwaardelijk voor een succesvolle toepassing van de richtlijn.

De Richtlijn passend bewijs preventie biedt een gemeenschappelijk kader om de effectiviteit van preventiebeleid overtuigend vast te stellen. De richtlijn helpt te bepalen welk bewijs passend is en benadrukt daarbij het geloofwaardig vaststellen van effecten, de toepasbaarheid in de praktijk en het in beeld brengen van de langetermijneffecten. Door een duidelijke minimumeis aan het bewijs over de werkzaamheid van preventie sluit de richtlijn aan bij wat wetenschappelijk mogelijk en maatschappelijk gewenst is om vast te stellen in hoeverre een preventiemaatregel voldoende onderbouwd is. Zo draagt de richtlijn bij aan een stevig fundament voor investeringen in preventie en vergroot het de kans op duurzame inbedding in beleid en financiering.

Richtlijn

Dit rapport benadrukt dat het versterken van preventiebeleid vraagt om stevig en passend bewijs. De richtlijn biedt daarvoor handvatten, met aandacht voor drie centrale pijlers: interne validiteit, toepasbaarheid en langetermijneffecten.

Interne validiteit gaat over de vraag of we werkelijk meten wat de interventie teweegbrengt. Het bewijs moet gebaseerd zijn op een overtuigende benadering van de “tegenfeitelijke situatie”: wat zou er gebeurd zijn zonder de maatregel? Hiervoor bestaan verschillende bewijsmethoden, variërend van grootschalige experimenten tot zorgvuldig uitgevoerde vergelijkingen met controlegroepen. De richtlijn stelt daarbij een minimale ondergrens gebaseerd op standaarden die op verschillende onderzoeksterreinen worden toegepast. Voor ingrijpende of risicovolle maatregelen kan een sterker bewijsniveau vereist zijn. De voorbeelden in dit rapport laten zien dat het in de praktijk mogelijk is aan deze standaard te voldoen.

De toepasbaarheid van resultaten is belangrijk omdat een maatregel in de ene context succesvol kan zijn, maar dat betekent niet automatisch dat dit elders ook zo is. De effectiviteit kan ook afhankelijk zijn van complementair preventiebeleid. De richtlijn stelt dat bij de beoordeling van bewijs goed moet worden gekeken naar de overeenkomsten en verschillen in context. Zeker bij complexe interventies is het belangrijk om ook de onderliggende werking en de voorwaarden voor succes in beeld te hebben. Alleen dan kan worden vastgesteld of de resultaten bruikbaar zijn voor beleid en uitvoering in de gewenste omgeving. Het is belangrijk om te vermelden dat de wetenschappelijke standaard die voor interne validiteit bestaat en in verschillende wetenschapsgebieden is uitgekristalliseerd voor toepasbaarheid niet beschikbaar is. De afweging zal daarom zorgvuldig gemaakt moeten worden langs de lijnen die de richtlijn hiervoor aandraagt.

Tot slot zijn langetermijneffecten van belang. Preventieve maatregelen hebben soms pas na jaren effect, terwijl studies meestal kortere termijnen bestrijken. Ook denken beleid en politiek dikwijls op een kortere termijn dan voor preventieve maatregelen gewenst is. De richtlijn stelt dat modellen en goed gekozen intermediaire uitkomsten, zoals veranderingen in leefstijl of gezondheidsgedrag, nodig zijn om deze effecten in te schatten. Dit vraagt om zorgvuldige onderbouwing en transparantie over onzekerheden die hiermee gepaard gaan. De richtlijn geeft concrete aanwijzingen hoe dit te doen.

Richting

Deze richtlijn markeert een belangrijke stap in het vaststellen van een gemeenschappelijk kader voor passend bewijs bij preventieve maatregelen. Daarmee wordt een richting gegeven, maar is niet het laatste woord gesproken. Preventie is een breed terrein waarin niet elke maatregel binnen de kaders van de richtlijn zal passen. Het is daarom van belang het principe 'pas toe of leg uit' te hanteren: waar van de richtlijn wordt afgeweken, moet dit grondig gemotiveerd en uitermate transparant gebeuren.

Tegelijkertijd moet de richtlijn innovatie en experiment niet in de weg staan. Nieuwe interventies verdienen de kans om in de praktijk te worden beproefd, ook als het beschikbare bewijs in eerste instantie nog beperkt is. Investeren in onderzoek dat passend bewijs oplevert, zal in het algemeen ook alleen gedaan worden als er al kleinschalige en soms ongecontroleerde studies over de interventie beschikbaar zijn. Juist door te experimenteren kan op termijn sterker bewijs worden opgebouwd over de effectiviteit van preventieve maatregelen.

De adviescommissie erkent ook dat urgente maatschappelijke vraagstukken niet altijd op passend bewijs kunnen wachten. Soms is preventiebeleid nodig uit voorzorg, ook al weten we nog niet precies wat de effecten ervan zijn. Het is aan de politiek om te beslissen wanneer preventiebeleid passend bewijs volgens deze richtlijn vereist en wanneer een minder stevige onderbouwing volstaat.

Het gebruik van de richtlijn vraagt bovendien om voortdurende monitoring en evaluatie. Het zou goed zijn over enkele jaren vast te stellen in hoeverre de richtlijn daadwerkelijk leidt tot een hogere kwaliteit van bewijs, of sommige interventies onvoldoende aan bod komen, en of aanpassing van de richtlijn nodig is. Daarnaast blijft het essentieel om te blijven investeren in onderzoek naar de effecten van preventie, zodat beleidskeuzes steeds beter onderbouwd en duurzaam verankerd kunnen worden.

Route

De richtlijn volgt op de studie van de Technische werkgroep kosten en baten van preventie (TWG-KBP, 2023). Deze werkgroep benadrukt dat beleidsmakers breder moeten kijken naar bewijs dan alleen naar gerandomiseerde studies. Preventie kent vaak maatregelen waarvoor gerandomiseerde experimenten niet haalbaar zijn; daarom is ook passend bewijs uit observationele data, intermediaire uitkomstmaten en modellering nodig. Deze richtlijn geeft hier invulling aan.

De richtlijn heeft dezelfde uitgangspunten als de werkwijzer voor MKBA in het sociaal domein en sluit op hoofdlijnen aan bij de benaderingen van ZIN en de Gezondheidsraad. De hogere bewijslast die de Gezondheidsraad hanteert voor bijvoorbeeld vaccinaties en bevolkingsonderzoeken sluit aan bij de richtlijn, omdat het gaat om interventies die tot gezondheidsschade kunnen leiden. Het valt niet uit te sluiten dat de benaderingen van deze twee instituten in sommige gevallen tot andere conclusies leiden. De adviescommissie beveelt aan om naar consistentie in de beoordeling van passend bewijs voor preventie toe te werken.

Het bewijs is op dit moment nog niet voor alle maatregelen op voldoende niveau. Als we de richtlijn serieus nemen, dan betekent het ook dat we moeten investeren in onderzoek om effecten op deze manier te onderbouwen. Wat de adviescommissie en de technische werkgroep adviseren is om de middelen voor de financiering van onderzoek naar potentieel waardevolle preventiemaatregelen zo in te zetten dat het bewijs ook daadwerkelijk passend is. Daarbij kan met de verdeling van onderzoeksgelden rekening worden gehouden met de kosten van het betrekken van een controlegroep, zodat gecorrigeerd wordt voor het feit dat dit bij sommige aanpakken moeilijker te realiseren is.

Een infrastructuur van observationele data om te monitoren en evalueren is hierbij van belang. Dit vergroot de mogelijkheden voor onderzoek op basis van niet-experimentele methoden dat van observationele data afhankelijk is. Wat hiervoor nodig is op het gebied van de verzameling en ontsluiting van gegevens ziet de adviescommissie als een belangrijke kennisvraag.

Tot besluit

Deze richtlijn is een belangrijke stap om investeringen in preventie goed onderbouwd op hun maatschappelijke waarde te kunnen beoordelen. Het doel van deze richtlijn is laten zien welk bewijs minimaal nodig is om preventiebeleid passend te onderbouwen. Uit een breed scala aan voorbeelden blijkt bovendien dat dit haalbaar is. Om de richtlijn effectief te laten zijn moet die een duidelijke plek krijgen in het beleid en moet er een partij komen die op naleving toeziet.

Bronnenlijst

- Agachi, E., Mierau, J. O., van Ittersum, K., & Bijmolt, T. H. (2024). The impact of a lifestyle behaviour change program on healthcare costs: Quasi-experimental real-world evidence from an open-access mobile health app in the Netherlands. *Preventive Medicine*, 189, 108174.
- Allender, S., Millar, L., Hovmand, P., Bell, C., Moodie, M., et al. (2020). Whole of systems trial of prevention strategies for childhood obesity: Implementation evaluation of the SEA Change Portland case study. *BMC Public Health*, 20, 117.
- Allender, S., Orellana, L., Crooks, N., Bolton, K.A., Fraser, P., et al. (2021). Four-Year Behavioral, Health-Related Quality of Life, and BMI Outcomes from a Cluster Randomized Whole of Systems Trial of Prevention Strategies for Childhood Obesity. *Obesity*, 29, 1022-1035.
- Altunkaya, J., Li, X., Adler, A., Feenstra, T., Fridhammar, A., et al. (2024). Examining the Impact of Structural Uncertainty Across 10 Type 2 Diabetes Models: Results From the 2022 Mount Hood Challenge. *Value in Health*, 27(10), 1338-1347.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512), 1331-1344.
- Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2025). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Review of Economic Studies*, rdaf087.
- van Baal, P. H. M., Polder, J. J., De Wit, G. A., Hoogenveen, R. T., Feenstra, T. L., et al. (2008). Lifetime medical costs of obesity: prevention no cure for increasing health expenditure. *PLoS medicine*, 5(2), e29.
- Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., & Sant'Anna, P. H. (2025). Difference-in-differences designs: A practitioner's guide. *arXiv preprint arXiv:2503.13323*.
- Bartelink, N. H., van Assema, P., Kremers, S. P., Savelberg, H. H., Oosterhoff, M., et al. (2019a). One- and two-year effects of the healthy primary school of the future on children's dietary and physical activity behaviours: A quasi-experimental study. *Nutrients*, 11(3), 689.
- Bartelink, N. H., Van Assema, P., Jansen, M., Savelberg, H., Moore, G. F., et al. (2019b). Process evaluation of the healthy primary School of the Future: the key learning points. *BMC Public Health*, 19(1), 698.
- Bartelink, N. H., van Assema, P., Jansen, M., Savelberg, H., & Kremers, S. (2019c). The moderating role of the school context on the effects of the healthy primary school of the future. *International Journal of Environmental Research and Public Health*, 16(13), 2432.
- Briggs, A. D., Mytton, O. T., Kehlbacher, A., Tiffin, R., Rayner, M., & Scarborough, P. (2013). Overall and income specific effect on prevalence of overweight and obesity of 20% sugar sweetened drink tax in UK: econometric and comparative risk assessment modelling study. *Bmj*, 347.
- Brunwasser, S. M., Gillham, J. E., & Kim, E. S. (2009). A meta-analytic review of the Penn Resiliency Program's effect on depressive symptoms. *Journal of consulting and clinical psychology*, 77(6), 1042.

- Burchett, H. E. D., Blanchard, L., Kneale, D., & Thomas, J. (2018). Assessing the applicability of public health intervention evaluations from one setting to another: a methodological study of the usability and usefulness of assessment tools and frameworks. *Health research policy and systems*, 16(1), 88.
- Busch, V., Steenkamer, I., van Nassau, F., van Opdorp, P., van Houtum, L., et al. (2024). The Effects of the Jump-In Whole-School Intervention on the Weight Development of Children in Amsterdam, the Netherlands. *Journal of School Health*, 94(1), 37-46.
- Buyse, M., & Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 1014-1029.
- Caro, J. J., Eddy, D. M., Kan, H., Kaltz, C., Patel, B., et al. (2014). Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value in health*, 17(2), 174-182.
- Carpenter, C., & Dobkin, C. (2011). The minimum legal drinking age and public health. *Journal of Economic Perspectives*, 25(2), 133-156.
- Cavalcanti, D. M., Ordoñez, J. A., da Silva, A. F., Basterra, E. L., Moncayo, A. L., et al. (2025). Health effects of the Brazilian Conditional Cash Transfer programme over 20 years and projections to 2030: a retrospective analysis and modelling study. *The Lancet Public Health*, 10(7), e548 - e558.
- CBS (2025a). Gezonde levensverwachting; vanaf 1981, Centraal Bureau voor de Statistiek.
- CBS (2025b). Uitgaven gezondheidszorg stegen in 2024 met 8,1 procent, Centraal Bureau voor de Statistiek.
- Chamberlain, R. C., Fecht, D., Davies, B., & Lavery, A. A. (2023). Health effects of low emission and congestion charging zones: a systematic review. *The Lancet public health*, 8(7), e559-e574.
- Chen, L., Appel, L. J., Loria, C., Lin, P. H., Champagne, C. M., et al. (2009). Reduction in consumption of sugar-sweetened beverages is associated with weight loss: the PREMIER trial. *The American Journal of Clinical Nutrition*, 89(5), 1299-1306.
- Ciani, O., Grigore, B., Blommestein, H., de Groot, S., Möllenkamp, M., et al. (2021). Validity of surrogate endpoints and their impact on coverage recommendations: a retrospective analysis across international health technology assessment agencies. *Medical Decision Making*, 41(4), 439-452.
- Coalitie Leefstijl in de Zorg (2025). Handreiking bewijsmethoden leefstijlinterventies in de zorg: Mogelijkheden voor het evalueren van effectiviteit.
- Commissie Theeuwes (2012). Durf te meten: Eindrapport Expertwerkgroep Effectmeting.
- Dawson, J., McCarthy, I., Taylor, C., Hildenbrand, K., Leamy, M., et al. (2021). Effectiveness of a group intervention to reduce the psychological distress of healthcare staff: a pre-post quasi-experimental evaluation. *BMC health services research*, 21(1), 392.
- Dickson, A., Gehrsitz, M., & Kemp, J. (2025). Does a spoonful of sugar levy help the calories go down? An analysis of the UK Soft Drinks Industry Levy. *Review of Economics and Statistics*, 1-10.
- van der Ende, M., Pees, R. O., Horlings, C., Batura, O., Wouterse, B., & Hoogendoorn-Lips, M. (2024). Suikerbelasting: Effectenonderzoek. Ecorys.
- Farrington, D.P., Gottfredson, D.C., Sherman, L.W., & Welsh, B.C. (2002). The Maryland scientific methods scale, *Evidence-based crime prevention*, 13-21.

FDA (2025). Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure. Food and Drug Administration.

Ferdinands, J. M., Blanton, L. H., Alyanak, E., Chung, J. R., Trujillo, L., et al. (2024). Protection against influenza hospitalizations from enhanced influenza vaccines among older adults: A systematic review and network meta-analysis. *Journal of the American Geriatrics Society*, 72(12), 3875-3889.

Finkelstein, A., Zhou, A., Taubman, S., & Doyle, J. (2020). Health care hotspotting—a randomized, controlled trial. *New England Journal of Medicine*, 382(2), 152-162.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in medicine*, 15(11), 1069-1092.

Gerritsen, S., Webbink, D., & Ter Weel, B. (2017). Sorting around the discontinuity threshold: The case of a neighbourhood investment programme. *De Economist*, 165(1), 101-128.

Goedendorp, M. M., Kuiper, D., Reijneveld, S. A., Sanderman, R., & Steverink, N. (2017). Sustaining program effectiveness after implementation: The case of the self-management of well-being group intervention for older adults. *Patient education and counseling*, 100(6), 1177-1184.

Goedendorp, M. M., & Steverink, N. (2017). Interventions based on self-management of well-being theory: pooling data to demonstrate mediation and ceiling effects, and to compare formats. *Aging & Mental Health*, 21(9), 947-953.

Gracner, T., Boone, C., & Gertler, P. J. (2024). Exposure to sugar rationing in the first 1000 days of life protected against chronic disease. *Science*, 386(6725), 1043-1048.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., et al. (2011). GRADE guidelines: 8. Rating the quality of evidence—indirectness. *Journal of clinical epidemiology*, 64(12), 1303-1310.

He, X., Sankaridurg, P., Wang, J., Chen, J., Naduvilath, T., et al. (2022). Time outdoors in reducing myopia: a school-based cluster randomized trial with objective monitoring of outdoor time and light intensity. *Ophthalmology*, 129(11), 1245-1254.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *Bmj*, 350.

Heymans, J. M., Kleijnen, S., & Verstijnen, I. M. (2013). Passend bewijs bij het bepalen van effectiviteit van interventies. *Nederlands Tijdschrift Voor Geneeskunde*, 157, A5479.

de Jonge-Heesen, K. W., Rasing, S. P., Vermulst, A. A., Scholte, R. H., Van Etteken, K. M., et al. (2020). Randomized control trial testing the effectiveness of implemented depression prevention in high-risk adolescents. *BMC medicine*, 18(1), 188.

Kim, K., Bretz, F., Cheung, Y. K. K., & Hampson, L. V. (Eds.). (2021). *Handbook of statistical methods for randomized controlled trials*. CRC Press.

de Koning, H. J., van Der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., et al. (2020). Reduced lung-cancer mortality with volume CT screening in a randomized trial. *New England journal of medicine*, 382(6), 503-513.

Koning, P., Webbink, D., & Koopmans, C. (2013). Onderwijs en arbeidsmarkt in MKBA's. In *Twee notities voor de algemene leidraad voor maatschappelijke kostenbatenanalyse* (Discussion Paper 73). SEO Economisch Onderzoek.

- Koopmans, C., Heyma, A., Hof, B., Imandt, M., Kok, L., & Pomp, M. (2016). Werkwijzer voor kosten-batenanalyse in het sociale domein. SEO Economisch Onderzoek.
- Koopmans, C., van Eijkel, R., & de Swart, L. (2024). Maatschappelijke kosten-batenanalyse en het beleidskompas: Past dat bij elkaar? SEO Economisch Onderzoek.
- Kopec, J. A., Finès, P., Manuel, D. G., Buckeridge, D. L., Flanagan, W. M., et al. (2010). Validation of population-based disease simulation models: a review of concepts and methods. *BMC Public Health*, 10(1), 710.
- Kremers, I., Steverink, N., Albersnagel, F. & Slaets, J. (2006). Improved selfmanagement ability and well-being in older women after a short group intervention. *Aging and Mental Health*, 10, 476-484.
- Lancaster, T., & Stead, L. F. (2017). Individual behavioural counselling for smoking cessation. *Cochrane database of systematic reviews*, (3).
- Lemp, J. M., Bommer, C., Xie, M., Michalik, F., Jani, A., et al. (2023). Quasi-experimental evaluation of a nationwide diabetes prevention programme. *Nature*, 624(7990), 138-144.
- Liu, E. F., Young, D. R., Sidell, M. A., Lee, C., Cohen, D. A., et al. (2025). City-Level Sugar-Sweetened Beverage Taxes and Changes in Adult Body Mass Index. *JAMA Network Open*, 8(1), e2456170-e2456170.
- Look AHEAD Research Group. (2010). Long-term effects of a lifestyle intervention on weight and cardiovascular risk factors in individuals with type 2 diabetes mellitus: four-year results of the Look AHEAD trial. *Archives of internal medicine*, 170(17), 1566-1575.
- Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1), 159-208.
- Margaryan, S. (2021). Low emission zones and population health. *Journal of Health Economics*, 76, 102402.
- McCaul, M., Naude, C., Neumann, I., Brennan, S., Meerpohl, J., et al. (2025). Outcomes. In *GRADE book version 1.0 (laatst geüpdatet augustus 2025)*. The GRADE Working Group.
- Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., et al. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *bmj*, 350.
- Moreno, E., Blasco, M., Porrón, C., Santana, G., Sardà, M., et al. (2025). VINCat program: Impact of a surveillance program on alcohol-based hand rub consumption in acute care facilities in Catalonia, Spain. *Enfermedades Infecciosas y Microbiología Clínica*, 43, S98-S105.
- Øvrebo, B., Halkjelsvik, T. B., Meisfjord, J. R., Bere, E., & Hart, R. K. (2020). The effects of an abrupt increase in taxes on candy and soda in Norway: an observational study of retail sales. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1), 115.
- Palmer, A. J., & Mount Hood 5 Modeling Group. (2013). Computer modeling of diabetes and its complications: a report on the Fifth Mount Hood challenge meeting. *Value in Health*, 16(4), 670-685.
- Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review-a new method of systematic review designed for complex policy interventions. *Journal of health services research & policy*, 10(1_suppl), 21-34.

- Pelletier, K. R. (2011). A review and analysis of the clinical and cost-effectiveness studies of comprehensive health promotion and disease management programs at the worksite: update VIII 2008 to 2010. *Journal of occupational and environmental medicine*, 53(11), 1310-1331.
- Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., et al. (2017). Making sense of complexity in context and implementation: the Context and Implementation of Complex Interventions (CICI) framework. *Implementation science*, 12(1), 21.
- Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4), 431-440.
- Rijksoverheid (2025). Zorgpartijen en ministerie van VWS sluiten zorgakkoord. Rijksoverheid.
- RIVM (2021). Preventief gezondheidsbeleid 2006-2018. Wat zijn de effecten en lessen? Rijksinstituut voor Volksgezondheid en Milieu.
- RIVM (2023). Impact van een getrapte verbruiksbelasting op de verkoop van suiker via alcoholvrije dranken. Rijksinstituut voor Volksgezondheid en Milieu.
- Romijn, G., & Renes, G. (2013). Algemene leidraad voor maatschappelijke kosten-batenanalyse. Centraal Planbureau & Planbureau voor de Leefomgeving.
- Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2), 2218-2244.
- Ruijsbroek, A., Wong, A., Kunst, A. E., van den Brink, C., van Oers, H. A., et al. (2017). The impact of urban regeneration programmes on health and health-related behaviour: Evaluation of the Dutch District Approach 6.5 years from the start. *PLoS One*, 12(5), e0177262.
- Samenwerkingsverband Erkenning van Interventies (2024). Kies erkend, krijg erkenning. Criteria 2024–2027. Samenwerkingsverband Erkenning van Interventies.
- Schloemer, T., & Schröder-Bäck, P. (2018). Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. *Implementation Science*, 13(1), 88.
- Schünemann, H. J., Brožek, J., Guyatt, G., & Oxman, A. (2013). GRADE Handbook. The GRADE Working Group.
- Schünemann, H. J., Neumann, I., Brennan, S., Meerpohl, J., Langendam, M., et al. (2025). Indirectness. In *GRADE book version 1.0 (laatst geüpdatet augustus 2025)*. The GRADE Working Group.
- Sevak, P., Stepanczuk, C. N., Bradley, K. W., Day, T., Peterson, G., et al. (2018). Effects of a community-based care management model for super-utilizers. *Am J Manag Care*, 24(11), e365-e370.
- Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., et al. (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *bmj*, 374(2061).
- Song, Z., & Baicker, K. (2019). Effect of a workplace wellness program on employee health and economic outcomes: a randomized clinical trial. *Jama*, 321(15), 1491-1501.
- Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3), 557-586.

- Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., et al. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *bmj*, 366.
- Steverink, N. (2009). Self-management ability scale: SMAS-30/versie 2. *Achtergrond, handleiding en scoring*.
- Steverink, N. (2014). Successful development and aging: Theory and intervention. In *The Oxford Handbook of Geropsychology*. Oxford University Press.
- Thomas, J., Petticrew, M., Noyes, J., Chandler, J., Rehfuess, et al. (2019). Chapter 17: Intervention complexity. In *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 451-477). Cochrane.
- Thomas, K. H., Martin, R. M., Davies, N. M., Metcalfe, C., Windmeijer, F., & Gunnell, D. (2013). Smoking cessation treatment and risk of depression, suicide, and self harm in the Clinical Practice Research Datalink: prospective cohort study. *Bmj*, 347.
- Torgerson, D. J. (2001). Contamination in trials: is cluster randomisation the answer? *Bmj*, 322(7282), 355-357.
- TWG-KBP (2023). Preventie op waarde schatten. Advies technische werkgroep kosten en baten van preventie.
- Vargas-Martínez, A. M., Trapero-Bertran, M., Lima-Serrano, M., Anokye, N., Pokhrel, S., & Mora, T. (2019). Measuring the effects on quality of life and alcohol consumption of a program to reduce binge drinking in Spanish adolescents. *Drug and Alcohol Dependence*, 205, 107597.
- Vemer, P., Corro Ramos, I., Van Voorn, G. A. K., Al, M. J., & Feenstra, T. L. (2016). AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. *Pharmacoeconomics*, 34(4), 349-361.
- van der Vliet, N., Suijkerbuijk, A. W., de Blaeij, A. T., de Wit, G. A., van Gils, et al. (2020). Ranking preventive interventions from different policy domains: what are the most cost-effective ways to improve public health?. *International journal of environmental research and public health*, 17(6), 2160.
- VWS (2025). Instellingsbesluit Adviescommissie richtlijn passend bewijs preventie. Ministerie van Volksgezondheid, Welzijn en Sport.
- Wagij, M. A., Crone, M., Bruinsma-van Zwicht, B., van Lith, J., Billings, D., & Rijnders, M. (2024). The Effect of CenteringPregnancy Group Antenatal Care on Maternal, Birth, and Neonatal Outcomes Among Low-Risk Women in the Netherlands: A Stepped-Wedge Cluster Randomized Trial. *Journal of midwifery & women's health*, 69(2), 191-201.
- Walker, V., Sanderson, E., Levin, M. G., Damrauer, S. M., Feeney, T., & Davies, N. M. (2024). Reading and conducting instrumental variable studies: guide, glossary, and checklist. *bmj*, 387.
- de Weerd, E., van Kippersluis, H., Zhang, L., & Roodbeen, R. (2024). *Economische Statistische Berichten*, 109(4839), 491-493.
- What Works Centre for Local Economic Growth (2016). Guide to scoring evidence using the Maryland Scientific Methods Scale. What Works Centre for Local Economic Growth.
- Wijnhoven, L. A., Creemers, D. H., Vermulst, A. A., Scholte, R. H., & Engels, R. C. (2014). Randomized controlled trial testing the effectiveness of a depression prevention program ('Op Volle Kracht')

among adolescent girls with elevated depressive symptoms. *Journal of abnormal child psychology*, 42(2), 217-228.

Willeboordse, M., Bartelink, N. H. M., van Assema, P., Kremers, S. P. J., Savelberg, H. H. C. M., et al. (2022). Battling the obesity epidemic with a school-based intervention: Long-term effects of a quasi-experimental study. *PLoS One*, 17(9), e0272291.

WRR (2021). Kiezen voor houdbare zorg. Mensen, middelen en maatschappelijk draagvlak. Wetenschappelijke Raad voor het Regeringsbeleid.

Wu, P. C., Chen, C. T., Lin, K. K., Sun, C. C., Kuo, C. N., et al. (2018). Myopia prevention and outdoor light intensity in a school-based cluster randomized trial. *Ophthalmology*, 125(8), 1239-1250.

Zorginstituut Nederland (2023a). Rapport - Pakketbeheer in de Praktijk 4. Zorginstituut Nederland.

Zorginstituut Nederland (2023b). Beoordeling stand van de wetenschap en praktijk 2023. Zorginstituut Nederland.

Zorginstituut Nederland (2024a). Rapport - Beoordelingskader kosteneffectiviteit van zorg. Zorginstituut Nederland.

Zorginstituut Nederland (2024b). Onzekerheid en Value of Information analyses. Verdiepingsmodule. Zorginstituut Nederland.

Zorginstituut Nederland (2024c). Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. Zorginstituut Nederland.

Bijlage A Gesprekspartners

De adviescommissie heeft gesproken met de volgende personen (in alfabetische volgorde):

- Prof. dr. Pim van Assendelft (Radboudumc en NFU)
- Sebastiaan Baan MA. (GGD West-Brabant)
- Dr. Jolanda van Bilsen (TNO)
- Marijke Bos MSc. (PGGM)
- Prof. dr. Marcel Canoy (Vilans en Vrije Universiteit Amsterdam)
- Dr. Afke Donker (Nji)
- Dr. Anoukh van Giessen (Universiteit Twente)
- Dr. Karin Hek (Nivel)
- Drs. Tom van 't Hek (Sportraad)
- Karen den Hertog MA. (GGD Amsterdam)
- Drs. Bert Hof (PBL)
- Dr. Stefanie Hofstede (FMS)
- Prof. dr. Xander Koolman (Vrije Universiteit Amsterdam)
- Lennart Langbroek MSc. (Sportraad)
- Prof. dr. Frank van Lenthe (Erasmus MC en NFU)
- Dr. Willem Jan Meerding (Algemene Rekenkamer)
- Dr. Madelon Meijer (NCJ)
- Annemiek de Nooijer MSc. (Andersson Elffers Felix en KAMG)
- Prof. dr. Johan Polder (RIVM en Tilburg University)
- Liliane de Ruiters MSc. (NFU)
- Prof. dr. Frederieke Schaafsma (Amsterdam UMC en KAMG)
- Dr. Caroline Schlinkert (Nivel)
- Ike van der Schoor BSc. (Nji)
- Prof. dr. Anja Schreijer (PDPC)
- Astrid Schulting MA. (GGD IJsselmeer)
- Dr. Ingrid Staal (NCJ)
- Dr. Foekje Stelma (Nivel)
- Dr. Leon van den Toorn (Erasmus MC en FMS)
- Ir. Edwin Velzel (PGGM)
- Prof. dr. Robert Verheij (Nivel en Tilburg University)
- Emma Verspoor MA. (Nji)
- Prof. dr. Robert Vonk (RVS en Universiteit Utrecht)

De adviescommissie heeft de opzet van de richtlijn gepresenteerd op een bijeenkomst met de volgende experts (in alfabetische volgorde):

- Prof. dr. Pieter van Baal (ESHPM)
- Prof. dr. Cornelis Boersma (Open Universiteit en UMCG)
- Prof. dr. Patrick Bossuyt (Amsterdam UMC)
- Prof. dr. Leona Hakkaart-van Roijen (ESHPM)
- Prof. dr. Anton Kunst (Amsterdam UMC)
- Prof. dr. ir. Jaap Seidell (emeritus hoogleraar Vrije Universiteit Amsterdam)
- Dr. Ben Wijnen (Trimbos-instituut)

Bijlage B Voorbeelden

1. Samen Gezond

Agachi, E., Mierau, J. O., van Ittersum, K., & Bijmolt, T. H. (2024). The impact of a lifestyle behaviour change program on healthcare costs: Quasi-experimental real-world evidence from an open-access mobile health app in the Netherlands. *Preventive Medicine*, 189, 108174.

PICOT

P	Volwassenen in Nederland die gebruikmaken van een open-access leefstijl-app (SamenGezond).
I	Deelname aan een leefstijlprogramma via de app. De app biedt functies zoals activiteiten bijhouden, educatieve artikelen, persoonlijke gezondheidsdoelen, quizen, online coaching en een dagelijkse 'fit-score'. Via interactieve elementen als chat, challenges en gepersonaliseerde doelen kunnen gebruikers punten verdienen, in te wisselen voor vouchers, kortingen, gadgets of donaties aan goede doelen
C	Niet-gebruikers van de app.
O	Gezondheidszorgkosten (huisartsenzorg, medisch-specialistische zorg).
T	De uitkomsten worden gemeten na twee jaar. De studie richt zich op einduitkomsten.

Resultaten

De zorgkosten daalden met 4,9% in het eerste jaar en 5,3% in het tweede jaar (ongeveer 154 euro per gebruiker per jaar), vooral door besparingen in huisarts- en specialistenzorg. Het effect was groter bij frequentere gebruikers. Hoewel het gebruik van geestelijke gezondheidszorg toenam, leidde dit niet tot hogere totale kosten en wijst dit mogelijk op vroege preventie. Er zijn geen negatieve neveneffecten gevonden.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences + propensity score matching
Sterke punten	- Groot databestand (n=15.506) - Voldoet aan parallelle-trends aanname - Vergelijkbare kenmerken van deelnemers in interventie- en controlegroep na matching
Zwakke punten	Mogelijke restconfounding (motivatie, gezondheidsgedrag), omdat alleen voor observeerbare kenmerken kan worden gecontroleerd.

Toepasbaarheid

Context: De studie vindt plaats in een real-world setting en is breed inzetbaar. Er worden bovendien diverse bevolkingsgroepen bereikt. De context is wel gebonden aan het Nederlandse zorgstelsel en de effectiviteit kan variëren bij andere apps of zorgsystemen.

Werkzame elementen: De studie bevat een korte beleidstheorie: App stimuleert gedragsverandering → minder zorggebruik → kostenbesparing. Er is echter geen procesevaluatie gedaan.

Zijn er andere studies? Ja, de bevindingen sluiten aan bij eerder bewijs dat mHealth-apps gedragsverandering kunnen bevorderen.

2. Begeleiding bij stoppen met roken

Lancaster, T., & Stead, L. F. (2017). Individual behavioural counselling for smoking cessation. *Cochrane database of systematic reviews*, (3).

PICOT

P	Rokers
I	• Begeleiding bij stoppen met roken: individuele begeleiding/counseling met specifieke componenten, gegeven door een getrainde zorgprofessional, buiten de gebruikelijke medische zorg. • Begeleiding zoals hierboven omschreven + ondersteunende farmacotherapie
C	Geen begeleiding, wel kort advies en/of zelfhulpmaterialen.
O	Gestopt zijn met roken.
T	Uitkomsten zijn gemeten op verschillende momenten, minimaal 6 maanden na start van de begeleiding. De studie richt zich op intermediaire uitkomsten, geen uiteindelijke uitkomsten. De relatie tussen roken en gezondheid is al veel onderzocht.

Resultaten

Individuele begeleiding verhoogt de kans op langdurig stoppen met roken gemiddeld met een risico-ratio van 1,57. Dit betekent dat waar bij een minimale interventie ongeveer 7 van de 100 personen zouden stoppen, dit aantal met begeleiding stijgt naar 10–12 per 100. Wanneer begeleiding wordt gecombineerd met ondersteunende farmacotherapie is er een klein extra effect, waardoor het aantal stoppers toeneemt van gemiddeld 11 naar 13 per 100 personen. Er zijn geen negatieve bijwerkingen of neveneffecten gerapporteerd.

Bewijsmethode

SMS-score	5
Methode	Systematische review van RCT's en natuurlijke experimenten.
Sterke punten	• Groot aantal studies van hoge kwaliteit geïncludeerd. Voor de beoordeling is gebruikgemaakt van het Cochrane 'Risk of bias'-instrument en de GRADE-methodiek.
Zwakke punten	• Grote heterogeniteit in de uitkomsten van studies die het effect van begeleiding bij stoppen met roken zonder farmacotherapie onderzoeken.

Toepasbaarheid

Context: Toepasbaarheid is vergroot door het grote aantal studies, met een hoog aantal deelnemers uit verschillende populaties (gezonde personen; patiënten in eerste of tweede lijn; verschillende bevolkingsgroepen, verschillen in motivatie om te stoppen met roken, e.a.) en uit uiteenlopende landen en werelddelen. Mogelijk kunnen factoren als gezondheidsszorgsysteem, gewoontes en beleid in een land e.a. variëren en toegang tot en uitvoering van deze interventie beïnvloeden.

Werkzame elementen: De studie bevat geen analyse van welke elementen uit de interventie werkzaam zijn. Dat stoppen met roken een positief gezondheidseffect heeft is echter meermaals aangetoond in eerdere studies.

Zijn er andere studies? Ja, in de systematische review wordt gebruikgemaakt van meerdere studies.

3. Cash transfers Brazilië

Cavalcanti, D. M., Ordoñez, J. A., da Silva, A. F., Basterra, E. L., Moncayo, A. L., et al. (2025). Health effects of the Brazilian Conditional Cash Transfer programme over 20 years and projections to 2030: a retrospective analysis and modelling study. *The Lancet Public Health*, 10(7), e548 - e558.

PICOT

P	Braziliaanse huishoudens met een laag inkomen in gemeentes met een hoge dekking van het programma (>30%)
I	Deelname aan het <i>Bolsa Família</i> Conditional Cash Transfer (CCT) programma en waar de dekking hoog was. Gemiddeld krijgen deelnemende huishoudens \$139 extra per maand.
C	Braziliaanse huishoudens in gemeentes met een lage dekking van het programma (<30%).
O	• Sterfte (totaal en leeftijdsspecifiek, o.a. onder-5 mortaliteit, mortaliteit bij ouderen) • Ziekenhuisopnames (totaal en leeftijdsspecifiek)
T	Uitkomsten zijn gemeten op verschillende momenten tussen 2000 en 2019. De gezondheidswinst wordt geprojecteerd tot 2030. De studie richt zich op einduitkomsten.

Resultaten

De studie toont aan dat hoge dekking en adequate uitkeringen sterk samenhangen met verbeterde gezondheidswinsten. Gedurende de eerste twintig jaar zijn door het programma naar schatting 713 duizend sterfgevallen en 8,2 miljoen ziekenhuisopnames voorkomen. De grootste effecten werden gevonden bij kinderen onder de vijf jaar (sterftereductie van circa 33%) en ouderen boven de 70 (bijna halvering van ziekenhuisopnames). Modelprojecties suggereren dat uitbreiding tot 2030 nog eens 684 duizend sterfgevallen en 8 miljoen opnames kan vermijden. Het onderzoek bevestigt dat sociale bescherming via CCT's niet enkel armoede vermindert, maar ook substantiële gezondheidswinst oplevert.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences + propensity score matching
Sterke punten	• De observeerbare kenmerken van de interventie- en controlegroep zijn vergelijkbaar, door propensity score matching.
Zwakke punten	• De parallelle-trends aanname wordt niet getoetst. • Het is niet duidelijk waarom de dekking in sommige gemeenten lager is. Hierdoor is het moeilijk om te beoordelen of deze gemeenten een adequate controlegroep zijn. De gevonden verbanden kunnen om die reden vertekend zijn (confounding door andere programma's bijvoorbeeld, of nationale trends in de economie).

Toepasbaarheid

Context: De grootte van het effect van een dergelijk sociaalzekerheidsprogramma verschilt vermoedelijk per land, afhankelijk van baseline-gezondheidszorg, armoedeniveau en mate van ongelijkheid. In dat opzicht is de externe generaliseerbaarheid naar hoge-inkomenslanden vermoedelijk beperkt. Aan de andere kant: er is geen reden om aan te nemen dat de algemene bevinding dat vermindering van armoede de gezondheid verbetert, geen algemene geldigheid heeft.

Werkzame elementen: De studie bevat een korte beleidstheorie: Conditional Cash Transfers → minder armoede → betere toegang tot zorg en betere leefomstandigheden → verbeterde gezondheid. Het onderzoek brengt daarnaast in kaart wat het relatieve effect is van de dekking en hoogte van de uitkering.

Zijn er andere studies? Ja, er zijn eerdere studies naar dezelfde interventie en verschillende studies naar andere CCT's.

4. Groepsinterventie tegen stress in de gezondheidssector

Dawson, J., McCarthy, I., Taylor, C., Hildenbrand, K., Leamy, M., et al. (2021). Effectiveness of a group intervention to reduce the psychological distress of healthcare staff: a pre-post quasi-experimental evaluation. *BMC health services research*, 21(1), 392.

PICOT

P	Personeel van gezondheidsorganisaties in het VK (8 ziekenhuizen, een ggz en een hospice).
I	'Rounds', een organisatiebrede groepsinterventie. De bijeenkomsten worden maandelijks gehouden, meestal tijdens de lunch met 30–100+ deelnemers. Deelname is vrijwillig. Een bijeenkomst duurt een uur, start met een multidisciplinair panel dat een patiëntencasus of thema presenteert en ervaringen deelt met nadruk op sociale, psychologische en emotionele aspecten. Daarna volgt een groepsdiscussie onder begeleiding van een senior clinicus en een getrainde facilitator.
C	Niet-deelnemers aan Rounds-bijeenkomsten
O	• Psychische belasting • Werkbetrokkenheid • Compassie • Zelfreflectie
T	Uitkomsten zijn gemeten na 8 maanden. De studie richt zich op intermediaire uitkomsten, geen einduitkomsten.

Resultaten

Regelmatige deelnemers woonden gemiddeld vier Rounds-bijeenkomsten bij. Zij lieten een significant grotere afname in psychische belasting zien dan niet-deelnemers. Het onderzoek vindt geen significante effecten op werkbetrokkenheid, compassie en zelfreflectie.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences
Sterke punten	• Analyse controleert voor verschillen in observeerbare karakteristieken tussen de interventie- en controlegroep.
Zwakke punten	• Mogelijke restconfounding van niet-observeerbare verschillen tussen de interventie- en controlegroep. • De parallelle-trends aanname wordt niet getoetst.

Toepasbaarheid

Context: De effectiviteit van dergelijke interventies hangt waarschijnlijk samen met de beschikbaarheid van andere anti-stressprogramma's en de gemiddelde ervaren werkdruk door gezondheidspersoneel. In het Verenigd Koninkrijk is Rounds al een breed geïmplementeerde interventie. Omdat de studie gebruikmaakt van daadwerkelijke deelname aan de bijeenkomsten op locaties waar Rounds al langer gangbaar is, is er geen reden om aan te nemen dat deelname hoger of lager zou zijn wanneer dit op (nog) grotere schaal geïmplementeerd zou worden.

Werkzame elementen: De studie bevat een korte beleidstheorie, maar onderzoekt verder niet wat de werkzame elementen van de interventie zijn: Deelname aan Rounds → Werkbetrokkenheid, compassie en zelfreflectie → Minder psychische belasting.

Zijn er andere studies? Voor zover bekend zijn er geen andere effectmetingen gedaan van de Rounds-interventie. Er is wel eerder onderzoek gedaan naar andere groepsinterventies voor het verminderen van stress bij gezondheidsmedewerkers.

5. Griepvaccinatie

Ferdinands, J. M., Blanton, L. H., Alyanak, E., Chung, J. R., Trujillo, L., et al. (2024). Protection against influenza hospitalizations from enhanced influenza vaccines among older adults: A systematic review and network meta-analysis. *Journal of the American Geriatrics Society*, 72(12), 3875-3889.

PICOT

P	Oudere volwassenen (≥65 jaar) die in aanmerking komen voor jaarlijkse influenzavaccinatie.
I	Verschillende “enhanced” influenzavaccins, waaronder: 1. Hooggedoseerd trivalent/quadrivalent vaccin (HD-IIV) 2. Recombinant influenza vaccin (RIV4) 3. Adjuvanted influenza vaccin (aIIV)
C	• Standaarddosering geïnactiveerd influenzavaccin (SD-IIV) • Andere enhanced vaccins (via netwerk-meta-analyse)
O	Bescherming tegen influenza-gerelateerde ziekenhuisopnames
T	Uitkomsten zijn gemeten over een influenzaseizoen. De studie richt zich op einduitkomsten.

Resultaten

De resultaten toonden dat enhanced vaccins consequent superieur waren aan standaarddosering: met name HD-IIV liet een significante reductie in ziekenhuisopnames zien, gevolgd door aIIV en RIV4. Hoewel de relatieve verschillen tussen de enhanced vaccins onderling klein waren, bevestigde de analyse dat zij ouderen betere bescherming bieden.

Bewijsmethode

SMS-score	3+
Methode	Systematische review van RCT's en observationele studies.
Sterke punten	• Groot aantal studies geïncludeerd. • Voor de beoordeling is gebruikgemaakt van het Cochrane 'Risk of bias'-instrument.
Zwakke punten	• Merendeel van geïncludeerde studies is niet causaal. • De geïncludeerde RCT's lieten geen significant effect zien.

Toepasbaarheid

Context: Wanneer de omstandigheden in landen vergelijkbaar zijn (toegang tot zorg bijvoorbeeld), gelden de uitkomsten van deze studie waarschijnlijk ook voor andere (hoge-inkomens)landen.

Werkzame elementen: In het artikel worden de elementen van die omstandigheden die belangrijk zijn verder niet onderzocht of gespecificeerd.

Zijn er andere studies? Ja, in de systematische review wordt gebruikgemaakt van meerdere studies.

6. Bevolkingsonderzoek longkanker

de Koning, H. J., van Der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., et al. (2020). Reduced lung-cancer mortality with volume CT screening in a randomized trial. *New England journal of medicine*, 382(6), 503-513.

PICOT

P	Mannen (in primaire analyse) en een kleinere groep vrouwen (subgroepanalyses), 50-74 jaar oud, huidige of voormalige rokers, hoge rookbelasting, uit Nederland en België.
I	Volume-based, lage-dosis CT-screening (low-dose CT, zonder contrast), vier screeningrondes: bij start (T0), na 1 jaar, na 3 jaar en na 5,5 jaar.
C	Geen interventie.
O	Sterfte aan longkanker
T	Uitkomsten zijn gemeten tot 10 jaar na de interventie. Het onderzoek kijkt zowel naar intermediaire uitkomsten als naar einduitkomsten.

Resultaten

Deze trial (naam: NELSON) toont aan dat volume-based, lage-dosis CT-screening bij de groep hoogrisicorokers/gewezen rokers tot een significante vermindering van de sterfte aan longkanker leidt bij mannen na 10 jaar, met relatief lage percentages valse positieven en follow-upprocedures. Voor vrouwen zijn de resultaten veelbelovend, maar met grotere onzekerheid. Er is geen duidelijk verschil in overalmortaliteit (ongeacht oorzaak). Screening leidt tot vroegere diagnose en daarmee betere kansen voor genezing.

Bewijsmethode

SMS-score	5
Methode	RCT
Sterke punten	• Lange follow-upduur. • Gebruik van registratiedata. • Geen verschillen in karakteristieken tussen de interventie- en controlegroep.
Zwakke punten	• Kleine omvang groep vrouwen. • Risico op overdiagnose.

Toepasbaarheid

Context: De keuze voor hoogrisicorokers/gewezen rokers van 50-74 jaar komt overeen met die in screeningprogramma's wereldwijd. Dus relevant voor velen die in vergelijkbare demografische en risicoprofielen zitten. In landen met minder goed ontwikkelde zorgsystemen, andere deelnamepercentages, andere rokersprofielen, afwijkende bevolkingssamenstelling of verschillende screeningkosten en -acceptatie, kan de balans tussen voordelen en nadelen anders uitvallen.

Werkzame elementen: De studie bevat een korte beleidstheorie: Bevolkingsonderzoek → Eerdere diagnoses van longkanker → Minder sterfte aan longkanker.

Zijn er andere studies? Ja, in de Verenigde Staten.

7. Op Volle Kracht, depressiepreventieprogramma

de Jonge-Heesen, K. W., Rasing, S. P., Vermulst, A. A., Scholte, R. H., Van Ettehoven, K. M., et al. (2020). Randomized control trial testing the effectiveness of implemented depression prevention in high-risk adolescents. *BMC medicine*, 18(1), 188.

PICOT

P	Middelbare scholieren (12-16 jaar) met verhoogd risico op depressie in Nederland
I	'Op Volle Kracht' (OVK 2.0), een preventieprogramma van 8 sessies, gebaseerd op cognitieve gedragstherapie.
C	• Psycho-educatie: brochure, en twee e-mails met tips om de stemming te verbeteren (o.a. aanmoedigen fysieke activiteit e.a.)
O	• Zelfgerapporteerde (door de jongere resp. ouders) symptomen van depressie. • Klinische beoordeling van depressie door psycholoog.
T	Uitkomsten zijn gemeten na 6 en 12 maanden. Het onderzoek kijkt naar intermediaire uitkomsten, geen einduitkomsten.

Resultaten

Een afname van zelfgerapporteerde depressieve symptomen in de interventiegroep in vergelijking met de controlegroep na 12 maanden. Er trad verbetering op bij 38,3% van de jongeren in de interventiegroep en bij 12,5% in de controlegroep. In de interventiegroep was bij 2,1% een verslechtering te zien, tegenover 8,3% in de controlegroep; bij 59,6% in de interventiegroep respectievelijk 79,2% in de controlegroep was er geen verandering. Er was geen verschil in gediagnosticeerde depressies na 6 maanden.

Bewijsmethode

SMS-score	5
Methode	Cluster-RCT
Sterke punten	• Uitkomsten zowel op basis van zelfrapportage als op basis van diagnostiek door professional. • Beperkte uitval gedurende experiment.
Zwakke punten	• Mogelijke selectieve deelname aan RCT. Slechts 27% van benaderde jongeren deed mee. Enkele verschillen in observeerbare kenmerken die mogelijk duiden op selectiviteit. • Kleine sample (n=110).

Toepasbaarheid

Context: De effectiviteit van OVK 2.0 is in deze studie onderzocht in real world omstandigheden. Hiermee is het aannemelijk dat de interventie ook toepasbaar is in de rest van Nederland, mits de benodigde omstandigheden met betrekking tot samenwerking aanwezig zijn dan wel gefaciliteerd worden. Door het risico op selectieve deelname zijn de resultaten mogelijk niet representatief voor alle doelgroepen. Bovendien ontbraken onder andere scholieren van wie de ouders geen toestemming gaven voor deelname, en scholieren die de Nederlandse taal onvoldoende machtig waren om deel te nemen.

Werkzame elementen: Theoretische werking wordt niet in deze studie genoemd. Cognitieve gedragstherapie, waarop OVK is gebaseerd, is wel een gangbare/gevalideerde psychologische behandelmethode.

Zijn er andere studies? Ja. 'Op Volle Kracht' is een Nederlandse adaptatie van een langer bestaand Engelstalig programma, Penn Resiliency Program (zie o.a. Brunwasser et al., 2009). De Nederlandse versie werd al eerder in een RCT onderzocht onder op depressie gescreende Nederlandse jongeren (Wijnhoven et al., 2014).

8. Grip en Glans

Kremers, I., Steverink, N., Albersnagel, F. & Slaets, J. (2006). Improved selfmanagement ability and well-being in older women after a short group intervention. *Aging and Mental Health*, 10, 476-484.

PICOT

P	50+-vrouwen (geen eindleeftijd) in Nederland met een verminderd subjectief welbevinden, waaronder gevoelens van eenzaamheid, en weinig eigen regie (dreigen te) ervaren, met kans op een negatieve spiraal.
I	De Grip & Glans interventie bestaat uit zes wekelijkse bijeenkomsten en een terugkombijeenkomst na drie maanden. Groepen bestaan uit acht tot twaalf deelnemers en worden geleid door één of twee getrainde G&G-docenten. Elke bijeenkomst omvat, naast informatie en uitleg, oefeningen over het werken met de GLANS-schijf die de vijf kerndomeinen van welbevinden weergeeft, en het toepassen van zelfmanagementvaardigheden. Deze interventie is uitgevoerd in drie verschillende varianten (groepsvariant, zelfhulpinterventie en met huisbezoeken). Deze varianten bevatten dezelfde theoretische principes.
C	Geen interventie. De controlegroepen vulden alleen een vragenlijst in.
O	• Zelfmanagementvaardigheid • Welbevinden • Eenzaamheid
T	Uitkomsten zijn direct na de interventie gemeten en na 6 maanden. De studie kijkt naar intermediaire uitkomsten. Er wordt niet echt onderbouwd wat deze uitkomsten betekenen voor de gezondheid op de lange termijn. Wel geven ze aan dat de intermediaire uitkomstmaten zoals zelfmanagementvaardigheden en welbevinden de ouderen helpen langer zelfstandig te blijven en daarmee een bijdrage leveren aan gezond en vitaal ouder worden.

Resultaten

Na afloop van de interventie en na 6 maanden zijn er significante verbeteringen te zien wat betreft zelfmanagementvaardigheden, -welbevinden en -eenzaamheid. De controlegroep liet geen veranderingen zien. De effecten bleven bestaan na 6 maanden maar waren niet meer significant ten opzichte van de controlegroep omdat deze controlegroep iets verbeterde, mogelijk door het invullen van de vragenlijsten.

Bewijsmethode

SMS-score	5
Methode	RCT
Sterke punten	• Geen significante verschillen in kenmerken tussen de interventie- en controlegroep bij de start van het experiment.
Zwakke punten	• Ongeveer 10% van de deelnemers viel uit gedurende het experiment, o.a. omdat zij niet dachten dat zij iets hadden aan de interventie.

Toepasbaarheid

Context: De effectiviteit van de interventie is in verschillende Nederlandse gemeenten aangetoond in deze studie en in Goedendorp et al. (2017).

Werkzame elementen: In de [beschrijving](#) van de erkende interventie wordt de theorie van Zelf-Management-van-Welbevinden beschreven. Deze theorie wordt in Goedendorp et al. (2017) getest. Er zijn daarnaast diverse procesevaluaties gedaan naar de werving, het actuele gebruik, programma-integriteit en -waardering en benodigde randvoorwaarden voor implementatie.

Zijn er andere studies? Ja. In Goedendorp et al. (2017) wordt de effectiviteit van de RCT naar de groepsvariant (Kremers, 2006) vergeleken met de effectiviteit van de interventie nadat deze geïmplementeerd is in de praktijk van zorg en welzijn, dus zonder inmenging van onderzoekers. Er werden geen significante verschillen gevonden tussen de RCT en de IMP in toename in zelfmanagementvaardigheden en -welbevinden.

9. Preventie van onnodig gebruik ziekenhuiszorg

Sevak, P., Stepanczuk, C. N., Bradley, K. W., Day, T., Peterson, et al. (2018). Effects of a community-based care management model for super-utilizers. *Am J Manag Care*, 24(11), e365-e370.

PICOT

P	De interventie richtte zich op kwetsbare, complexe patiënten met een uitzonderlijk hoog zorgverbruik en een grote behoefte aan gecoördineerde, geïntegreerde zorg buiten het ziekenhuis. Het betrof een community-based programma in samenwerking met een zorgverzekeraar en een lokaal zorgnetwerk op vier stedelijke locaties in de VS, San Diego (CA), Aurora (CO), Kansas City (MO) en Allentown (PA) in de VS, in de periode 2013-2015. Het programma staat ook bekend als het Camden-model.
I	Elk deelnemend gezin kreeg een zorgmanager (nurse care manager of social worker) toegewezen. Deze werkte nauw samen met de huisarts en andere zorgverleners. Naast de care manager bestond het team vaak uit een verpleegkundige, maatschappelijk werker en een community health worker. Het team bezocht patiënten thuis en hielp met medicatiebeheer, afspraken maken, vervoer regelen en zorgcoördinatie. De aanpak was nadrukkelijk proactief, dus niet wachtten tot patiënten zich melden op de SEH. Er werd expliciet gekeken naar sociale determinanten van gezondheid: huisvesting, inkomen, toegang tot voeding en psychische ondersteuning.
C	Geen interventie, gebruikelijke zorgpraktijk.
O	• Aantal ziekenhuisopnames • Aantal heropnames (binnen 30 dagen) • Aantal spoedeisende-hulpbezoeken • Totale zorguitgaven (Medicare-uitgaven per persoon per maand)
T	Uitkomsten zijn 3, 6, 9 en 12 maanden na interventie gemeten. De studie richt zich op einduitkomsten.

Resultaten

Alleen bij de 30-dagenheropnames was er een forse daling (37%) in de interventiegroep ten opzichte van de controlegroep. Dit effect was echter slechts marginaal significant ($p = 0,086$). Voor de andere uitkomstmaten waren de verschillen niet significant.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences + matching
Sterke punten	• Er is rekening gehouden met clustering van de standaardfouten. • Geen verschillen in observeerbare kenmerken tussen interventie- en controlegroep na matching.
Zwakke punten	• Geen formele test op parallelle trends. Alleen beschrijvend bewijs. • Mogelijke restconfounding door niet-observeerbare verschillen tussen interventie- en controlegroep.

Toepasbaarheid

Context: De auteurs geven aan dat algemene uitspraken over de effectiviteit van de interventie op basis van deze studie niet mogelijk is. Waarschijnlijk hangen de effecten ook af van de rest van het zorgsysteem.

Werkzame elementen: De studie bevat geen beleidstheorie of analyse van de werkzame elementen.

Zijn er andere studies? Ja. Een later uitgevoerde RCT voor dezelfde interventie (Finkelstein et al. 2020) vindt géén effect. De (enige) uitkomstmaat bij Finkelstein et al. (2020) is heropname binnen 180 dagen, niet 30 dagen.

10. Myopie-preventie door schoolprogramma met meer dag-/buitenlicht

Wu, P. C., Chen, C. T., Lin, K. K., Sun, C. C., Kuo, C. N., et al. (2018). Myopia prevention and outdoor light intensity in a school-based cluster randomized trial. *Ophthalmology*, 125(8), 1239-1250.

PICOT

P	Schoolkinderen, leerlingen klas 1 (gemiddeld 6,3 jaar oud) in Taiwan
I	ROCT711-programma, Recess Outside Classroom-programma, bestaande uit: pauzes buiten de klas; outdoor-schoolactiviteiten, daglicht-paspoort in de weekends, opdrachten om buiten te leren tijdens zomervakantie; educatie over ooggezondheid met adviezen waaronder een 30/10-regel voor myopie-preventie; sportadvies; advies voor dagelijks 120 minuten activiteiten buitenshuis.
C	Gebruikelijk schoolprogramma zonder extra stimulatie om buiten te spelen, alleen educatie over ooggezondheid met adviezen waaronder een 30/10-regel voor myopie-preventie; sportadvies; advies voor dagelijks 120 minuten activiteiten buitenshuis.
O	• Mate van myopie (bijziendheid) • Aslengte van oog • Progressie van de myopie
T	Uitkomsten zijn gemeten na 1 jaar. Het onderzoek kijkt naar einduitkomsten.

Resultaten

In de interventiegroep kwam significant minder toename van myopie en minder groei van het oog (minder toename van de aslengte) voor dan in de controlegroep en een 54% lager risico op snelle progressie van de myopie. Het beschermende effect gold zowel voor kinderen met als zonder myopie, vergeleken met de controlegroep.

Bewijsmethode

SMS-score	5
Methode	Cluster-RCT
Sterke punten	• Vergelijkbare karakteristieken tussen de interventie- en controlegroep.
Zwakke punten	• Aandeel kinderen met myopie in de interventiegroep was relatief laag, omdat kinderen die voor hun myopie behandeld werden geen onderdeel waren van de sample. • Een derde van de deelnemende scholen viel uit gedurende het experiment.

Toepasbaarheid

Context: De studie is uitgevoerd in Oost-Azië. Het is te verwachten dat de interventie, wanneer toegepast in westerse landen, vergelijkbare resultaten zal hebben, vanwege de pathofysiologie (te weinig daglicht is een bekende risicofactor voor de ontwikkeling van myopie).

Werkzame elementen: In de (medisch-wetenschappelijke) literatuur is het positieve effect van blootstelling aan buiten-/daglicht op preventie van myopie en progressie van myopie welbekend. Meer blootstelling aan buitenlicht remt de groei van het oog (aslengte) en progressie van myopie.

Zijn er andere studies? De studie is gedaan onder meerdere scholen, maar wel in één land. Een andere studie hiernaar is een cluster-RCT in Shanghai (He et al., 2022).

11. Minimumleeftijd alcoholconsumptie

Carpenter, C., & Dobkin, C. (2011). The minimum legal drinking age and public health. *Journal of Economic Perspectives*, 25(2), 133-156.

PICOT

P	Jongeren in de Verenigde Staten
I	Minimale wettelijke leeftijd voor alcoholgebruik van 21 jaar.
C	Geen alcohol mogen drinken volgens de wet.
O	• Sterfte (totaal en door zelfmoord, verkeersongevallen, alcoholvergiftiging, moord en overig) • Alcoholconsumptie
T	Uitkomsten zijn gemeten tot 2 jaar voor en na de 21 ^{ste} verjaardag. De studie brengt zowel het effect op intermediaire uitkomsten als op einduitkomsten in kaart.

Resultaten

De minimale wettelijke leeftijd voor alcoholgebruik heeft een significant positief effect op sterfte door zelfmoord, verkeersongevallen en alcoholvergiftiging rond de leeftijd van 21 jaar. Jongeren die net ouder zijn dan 21 jaar consumeren bovendien meer alcohol dan jongeren die nog net geen 21 jaar zijn.

Bewijsmethode

SMS-score	4
Methode	Regression discontinuity design
Sterke punten	• Geen andere discontinuïteiten bij de drempelwaarde. • Geen beïnvloeding mogelijk van het halen van de drempelwaarde, want drempelwaarde is een leeftijd.
Zwakke punten	• Een bedreiging voor de interne validiteit is dat een deel van de toename in negatieve gebeurtenissen na het bereiken van de drinkleeftijd voortkomt uit het leren verantwoord drinken. Hierdoor stijgt de sterfte vaak in de eerste maanden, ongeacht of de leeftijdsgrens 18, 21 of hoger is. Analyses die deze daling later als 'geredde levens' meetellen, overschatten daardoor het effect van de minimumleeftijd.

Toepasbaarheid

Context: Effecten zijn geschat voor jongeren van ongeveer 21 jaar oud. De effecten zijn mogelijk anders voor jongeren van andere leeftijden.

Werkzame elementen: De studie bevat een korte beleidstheorie: Minimumleeftijd voor alcoholconsumptie → Minder alcoholconsumptie voor jongeren jonger dan de minimumleeftijd → Minder zelfmoord, verkeersongevallen en alcoholvergiftiging.

Zijn er andere studies? Ja, de Weerd et al. (2024) onderzoeken het effect van verhoging van de minimumleeftijd in Nederland.

12. Interventie tegen bingedrinken onder middelbare scholieren

Vargas-Martínez, A. M., Trapero-Bertran, M., Lima-Serrano, M., Anokye, N., Pokhrel, S., & Mora, T. (2019). Measuring the effects on quality of life and alcohol consumption of a program to reduce binge drinking in Spanish adolescents. *Drug and Alcohol Dependence*, 205, 107597.

PICOT

P	Andalusische adolescenten (15-19 jaar) die op urban public high schools zitten.
I	ALERTA ALCOHOL, een web-based computer-tailored interventie bestaand uit 6 sessies met kennisoverdracht en simulaties van sociale situaties waarin alcohol gedronken wordt. De interventie wordt per school uitgerold maar jongeren doorlopen de 6 sessies individueel.
C	Geen interventie.
O	• Aantal bingedrinksessies • Kwaliteit van leven
T	Uitkomsten zijn 4 maanden na de start van de interventie gemeten. De studie brengt intermediaire uitkomsten in kaart, maar de relatie tussen gezondheid en bingedrinksessies is veelvuldig aangetoond.

Resultaten

Het aantal bingedrinkepisodes nam af op de interventiescholen. Het effect op kwaliteit van leven werd in de vier maanden observatieperiode niet zichtbaar (volgens de auteurs *nog* niet zichtbaar).

Bewijsmethode

SMS-score	5
Methode	Cluster-RCT
Sterke punten	• Beperkt risico op beïnvloeding van de controlegroep door de interventiegroep, vanwege clustering op schoolniveau.
Zwakke punten	• Hoge uitval tijdens het onderzoek: van de 1247 deelnemende adolescenten (baseline) hebben slechts 612 jongeren ook de eindmeting ingevuld.

Toepasbaarheid

Context: De studie heeft plaatsgevonden in Andalusië. De effectiviteit van de interventie hangt waarschijnlijk samen met de context, zoals de drinkcultuur en de beschikbaarheid van andere interventies om bingedrinken onder jongeren tegen te gaan.

Werkzame elementen: Hoewel kwaliteit van leven als uitkomstmaat gebruikt is, wordt er geen heel duidelijke rationale gegeven voor de relatie tussen bingedrinken en kwaliteit van leven. Hier ontbreekt dus een theorie. Voor wat betreft bingedrinkepisodes is hiervoor zoveel bewijs van de schadelijkheid dat dit een prima intermediaire uitkomstmaat is, omdat de relatie met de uiteindelijke gezondheidsuitkomst dus zeker vaststaat. Ook wordt een dosis-effectrelatie aangetoond: jongeren die meer sessies volgden hebben een grotere reductie van het alcoholgebruik en bingedrinkepisodes.

Zijn er andere studies? Ja, er is eerder onderzoek gedaan naar interventies tegen bingedrinken onder middelbare scholieren, hoewel de vormgeving van de interventies wel verschilt.

13. Suikertaks Noorwegen

Øvrebo, B., Halkjelsvik, T. B., Meisfjord, J. R., Bere, E., & Hart, R. K. (2020). The effects of an abrupt increase in taxes on candy and soda in Norway: an observational study of retail sales. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1), 115.

PICOT

P	Winkels in Noorwegen
I	Accijnsverhoging per 1 januari 2018: 80% toename voor snoepgoed en chocolade, 40% voor niet-alcoholische dranken.
C	Winkels in Noorwegen vóór 1 januari 2018.
O	Verkooppvolumes
T	Uitkomsten worden gemeten direct na en een half jaar na de interventie. De studie kijkt naar intermediaire uitkomsten, geen einduitkomsten.

Resultaten

Het paper vindt insignificante en inconsistente effecten. Met difference-in-discontinuities wordt een toename in de verkoop van snoep en een daling van de verkoop van frisdrank gevonden, en in de difference-in-differences juist andersom.

Bewijsmethode

SMS-score	3/4
Methode	Difference-in-discontinuities en (soort) difference-in-differences
Sterke punten	• Grote dataset • Standaardfouten gecorrigeerd voor clustering
Zwakke punten	• In de difference-in-differences wordt niet gebruikgemaakt van een controlegroep op hetzelfde moment in de tijd. Hierdoor is het niet uit te sluiten dat andere tijdstrends toch een effect hebben. • In de difference-in-discontinuities spelen er mogelijk nog andere discontinuïteiten op de drempelwaarde van 1 januari.

Toepasbaarheid

Context: De studie onderzoekt vrij grote prijseffecten, dat maakt ze relevanter dan bijvoorbeeld de suikertaks in Mexico. Tegelijk zijn de effecten context-specifiek (denk aan kopen over de grens en de hoge accijns op alcoholische dranken).

Werkzame elementen: De studie bevat een korte beleidstheorie: Hogere prijzen → Lagere consumptie. De studie geeft verder geen inzicht in de werkzame elementen die zouden kunnen helpen om te begrijpen of effecten ook in een andere context zouden optreden.

Zijn er andere studies? Een andere studie op (een subsample van) dezelfde data vond wel een sterke reductie in verkopen van chocolade en frisdrank.

14. Milieuzones

Margaryan, S. (2021). Low emission zones and population health. *Journal of Health Economics*, 76, 102402. <https://doi.org/10.1016/j.jhealeco.2020.102402>

PICOT

P	Inwoners van grote steden in Duitsland waarvoor registerdata over poliklinische en klinische zorg beschikbaar zijn.
I	Invoering van Low Emission Zones (LEZ's) die vervuilende voertuigen weren uit aangewezen stedelijke gebieden.
C	Steden/perioden zonder LEZ (gefaseerde invoering maakt vergelijking vóór/na en tussen steden mogelijk).
O	Aantal patiënten met cardiovasculaire aandoeningen
T	Uitkomsten worden op verschillende momenten gemeten. De studie kijkt zowel naar intermediaire uitkomsten als einduitkomsten.

Resultaten

Milieuzones leiden tot een daling van het aantal patiënten met cardiovasculaire aandoeningen met circa 2–3% waarbij het effect sterker is bij ouderen van 65 jaar en ouder. Dit komt doordat milieuzones de verkeersgerelateerde luchtvervuiling verminderen. Er zijn geen aanwijzingen gevonden voor schadelijke neveneffecten.

Bewijsmethode

SMS-score	3
Methode	(Systematische review van) stepped wedge difference-in-differences studies
Sterke punten	• Grote populatie. • Trends van interventie- en controlegroep zijn parallel voor invoering van de milieuzones.
Zwakke punten	• Mogelijke residual confounding als parallele trends vóór invoering niet volledig gelijk zijn. • Potentiële heterogeniteit tussen steden (beleid, naleving, omvang milieuzones) kan effectschattingen beïnvloeden. • Gezondheidsuitkomst is patiëntenaantal; ernst/incidentie vs. prevalentie onderscheiden is beperkt.

Toepasbaarheid

Context: Resultaten gelden overtuigend voor Duitse grote steden met vergelijkbare implementatie en handhaving. Overdraagbaarheid naar andere landen/steden hangt af van wagenpark, naleving, zone-ontwerp en luchtkwaliteitsniveaus.

Werkzame elementen: Studie bevat korte beleidstheorie: Milieuzones beperken vuile voertuigen → Lagere concentraties PM₁₀/NO₂ → minder cardiovasculaire morbiditeit. Dit mechanisme is ook elders aangetoond.

Zijn er andere studies? Ja. Een recente systematic review (Chamberlain et al., 2023) geeft een overzicht van de literatuur en laat zien dat er veel studies zijn die aantonen dat LEZ's een positief effect hebben op de gezondheid.

15. Gezonde Basisschool van de Toekomst

Bartelink, N. H., van Assema, P., Kremers, S. P., Savelberg, H. H., Oosterhoff, M., et al. (2019a). One-and two-year effects of the healthy primary school of the future on children's dietary and physical activity behaviours: A quasi-experimental study. *Nutrients*, 11(3), 689.

Willeboordse, M., Bartelink, N. H. M., van Assema, P., Kremers, S. P. J., Savelberg, H. H. C. M., et al. (2022). Battling the obesity epidemic with a school-based intervention: Long-term effects of a quasi-experimental study. *PLoS One*, 17(9), e0272291.

PICOT

P	Basisschoolleerlingen (groep 1-8) in de Parkstad-regio in Nederland
I	De GBT is een aanpak waarbij de schooldag wordt verlengd en de schoolcontext wordt aangepast om een gezonde leefstijl te bevorderen. Kernelementen zijn het aanbieden van een gezond tussendoortje tijdens de ochtendpauze en een gezonde schoollunch volgens de normen van het Voedingscentrum, gezondheidseducatie, bewegend leren gedurende de schooldag en professioneel begeleidde beweeg- en cultuuractiviteiten tijdens de tussenschoolse opvang.
C	Geen interventie.
O	• BMI • Buikomtrek • Voedingsgedrag • Lichamelijke activiteit • Algeheel welbevinden • Pesten • Tevredenheid van leerlingen en ouders
T	Uitkomsten zijn gemeten na 1, 2, 3 en 4 jaar. De studies kijken naar intermediaire uitkomsten zoals BMI, voedingsgedrag en beweeggedrag. Dit zijn beschermende factoren voor het ontstaan van o.a. obesitas, hart- en vaatziekten en diabetes.

Resultaten

Significant lagere BMI-score na 4 jaar. Deze bleven onveranderd in de interventiescholen maar namen toe bij de controlescholen. Hetzelfde geldt voor de buikomtrek. Het voedingsgedrag veranderde wel op korte termijn maar dit effect was op lange termijn niet meer zichtbaar. De lichamelijke activiteit veranderde niet significant. Van het voedingsgedrag bleef alleen de toename van waterconsumptie en zuivelconsumptie significant na vier jaar. De resultaten (hogere effect sizes) zijn beter naarmate de kinderen langer blootgesteld zijn aan de interventies.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences
Sterke punten	• De analyses controleren voor verschillen in observeerbare kenmerken tussen de interventie- en controlegroep.
Zwakke punten	• Deelname was vrijwillig waardoor er mogelijk restconfounding is van niet-observeerbare verschillen tussen de interventie- en controlegroep. • Mogelijke onderschatting van het effect, omdat kinderen gemiddeld al gezonder waren op deelnemende scholen.

Toepasbaarheid

Context: De studie is uitgevoerd in Nederland, maar wel maar in één regio van Nederland.

Werkzame elementen: De interventie is een erkende interventie en is voorzien van een theoretische onderbouwing, procesevaluatie (Bartelink et al., 2019b) en inzicht in werkzame elementen. Deze zijn geformuleerd op basis van de theoretische onderbouwing en de procesevaluatie. Er zijn verschillende studies over het effect en de uitvoering van de BST maar deze zijn wel allemaal gebaseerd op dezelfde onderzoekspopulatie. Er is ook gekeken naar het effect van de context op de uitkomsten (Bartelink et al.,

2019c). Gekeken is naar type leerlingen, inzet/gedrag van leerkrachten en organisatorische issues op school. De studie kan de gesuggereerde moderatie-effecten in de huidige opzet echter niet bewijzen vanwege het kleine aantal scholen. Het is plausibel dat een combinatie van de genoemde factoren samenhangt met succes van de interventie, maar hoe groot de rol is van elk van die factoren kan niet worden bepaald.

Zijn er andere studies? Er zijn verschillende onderzoeken gedaan naar de Gezonde Basisschool van de Toekomst, maar deze vinden allemaal plaats op dezelfde locaties. Er zijn daarnaast wel eerdere studies beschikbaar naar de effecten van andere gezondheidsinterventies op (basis)scholen.

16. Jump-In-schoolprogramma rondom bewegen en voeding

Busch, V., Steenkamer, I., van Nassau, F., van Opdorp, P., van Houtum, L., et al. (2024). The Effects of the Jump-In Whole-School Intervention on the Weight Development of Children in Amsterdam, the Netherlands. *Journal of School Health*, 94(1), 37-46.

PICOT

P	Kinderen op basisscholen in Amsterdam met (boven)gemiddeld veel kinderen met overgewicht of obesitas
I	Jump-in is een Amsterdams schoolbreed programma dat gezonde voeding en bewegen stimuleert op basisscholen (4–12 jaar). Het programma bestaat uit drie componenten: (1) Voeding (verschillende beleidselementen gericht op gezonde voeding, educatieve workshops voor ouders en kinderen en integratie van het gezondheidseducatieprogramma 'Smaaklessen'), (2) Beweging (monitoring en ondersteuning van lichamelijke activiteit en motorische ontwikkeling, extra beweegmomenten voor alle kinderen, minimaal twee keer per week tijdens gym en na schooltijd en ondersteuning voor kinderen die zich anders geen lidmaatschap kunnen veroorloven) en (3) actieve pauzes. Het programma ondersteunt scholen bij het invoeren en verankeren van meerdere interventiecomponenten gedurende drie schooljaren.
C	Kinderen op scholen in andere gemeenten.
O	• (Gestandaardiseerde) BMI • Wisselen naar een gezondere gewichtscategorie • Wisselen naar een gezond gewicht
T	Uitkomsten worden gemeten na 5 jaar. Er wordt gekeken naar intermediaire uitkomsten (BMI en gewichtscategorieën).

Resultaten

De volledig uitgevoerde interventie leidde tot een daling van de gestandaardiseerde BMI ten opzichte van de controlegroep. Ook was de kans groter dat kinderen in de loop van de tijd naar een gezondere gewichtscategorie gingen, al werd geen statistisch significante verschuiving naar een gezond gewicht gevonden.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences
Sterke punten	• Geen selectieve deelname van scholen of uitval gedurende looptijd van het onderzoek.
Zwakke punten	• Parallele-trends aanname wordt niet getoetst.

Toepasbaarheid

Context: De studie onderzoekt de effectiviteit van Jump-In in Amsterdam, op scholen met bovengemiddeld veel kinderen met overgewicht of obesitas. Het is aannemelijk dat de effecten vergelijkbaar zouden zijn als deze in andere Nederlandse gemeenten zou worden ingevoerd, indien dit ook scholen met bovengemiddeld veel kinderen met overgewicht of obesitas zijn.

Werkzame elementen: De studie bevat een korte beleidstheorie: Jump-in → Gezondere voedingskeuzes en meer beweging → Lagere BMI. De effecten worden wel gerelateerd aan de mate waarin Jump-in is geïmplementeerd, maar de werkzame elementen van Jump-in zijn niet in kaart gebracht.

Zijn er andere studies? Ja. Er zijn verschillende andere studies die de effectiviteit van Jump-In hebben onderzocht. Deze kijken wel allemaal naar Amsterdam.

17. Stedelijke vernieuwing

Ruijsbroek, A., Wong, A., Kunst, A. E., van den Brink, C., van Oers, et al. (2017). The impact of urban regeneration programmes on health and health-related behaviour: Evaluation of the Dutch District Approach 6.5 years from the start. *PLoS One*, 12(5), e0177262.

PICOT

P	Inwoners van de 40 meest achtergestelde wijken in Nederland
I	Wijkaanpak, een stedelijk vernieuwingsprogramma dat vanaf 2007 leefomstandigheden in de 40 meest achtergestelde wijken moest verbeteren. De interventies richtten zich op werk, onderwijs, huisvesting, de fysieke leefomgeving, veiligheid en sociale cohesie. De interventies waren in elke wijk afgestemd op de specifieke lokale behoeften. Hierdoor bestonden er grote verschillen in het aantal bereikte inwoners, de mate van gerealiseerde verandering en het aantal verschillende soorten activiteiten.
C	Vergelijkbare wijken zonder Wijkaanpak
O	• Algemene gezondheid • Mentale gezondheid • Fysieke activiteit • Roken • Overgewicht • Obesitas
T	Uitkomsten worden op verschillende momenten gemeten tot 6,5 jaar na start van de interventie. De studie meet einduitkomsten.

Resultaten

Veranderingen in gezondheid en aan gezondheid gerelateerd gedrag tussen de pre-interventie- en interventieperiode waren ongeveer even groot in de doelwijken als in de controledistricten. De DiD-schattingen van het effect waren inconsistent en statistisch niet significant, ook niet wanneer onderscheid werd gemaakt naar sekse of naar de intensiteit van de interventieprogramma's.

Bewijsmethode

SMS-score	3
Methode	Difference-in-differences
Sterke punten	• Grotendeels vergelijkbare kenmerken in de interventie- en controlegroep door matching.
Zwakke punten	• Parallele-trends aanname wordt niet formeel getoetst. • Risico dat uitkomsten worden beïnvloed door veranderingen in samenstelling van de interventie- en controlegroep over tijd. • Kans op restconfounding vanwege niet-willekeurige toewijzing aan de interventiegroep.

Toepasbaarheid

Context: De studie vond plaats in 40 Nederlandse achterstandswijken. Het is aannemelijk dat de resultaten vergelijkbaar zijn in andere wijken met een vergelijkbaar sociaal-economische samenstelling en leefbaarheid.

Werkzame elementen: De studie bevat een korte beleidstheorie: Wijkaanpak → Verbetering in sociale determinanten van gezondheid, zoals werk, huisvesting en sociale cohesie → Verbeterde gezondheid. De manier waarop per wijk de aanpak is vormgegeven, wordt niet in kaart gebracht, en dus zijn de werkzame elementen van de aanpak ook niet te identificeren.

Zijn er andere studies? Ja, die kijken naar de kortetermijneffecten van dezelfde interventie. Er zijn daarnaast eerder studies gedaan naar vergelijkbare interventies in het buitenland.

18. Workplace wellness-programma

Song, Z., & Baicker, K. (2019). Effect of a workplace wellness program on employee health and economic outcomes: a randomized clinical trial. *Jama*, 321(15), 1491-1501.

PICOT

P	Medewerkers van 160 vestigingen van een bepaald Amerikaans bedrijf (32.974 werknemers).
I	Workplace wellness-programma met modules over voeding, beweging en stressreductie waaruit medewerkers zelf konden kiezen.
C	Geen interventie.
O	• Zelf-gerapporteerd leefstijlgedrag • Klinische biometrie (BMI, bloeddruk, cholesterol) • Zorgkosten • Ziekteverzuim • Werkgerelateerde uitkomsten
T	Uitkomsten worden op verschillende momenten tot 18 maanden na start interventie gemeten. De studie kijkt zowel naar intermediaire uitkomsten als einduitkomsten.

Resultaten

Geen significante effecten op de meeste uitkomsten: BMI, bloeddruk, cholesterol, zorgkosten, ziekteverzuim of werkprestaties. Alleen positieve effecten voor zelf-gerapporteerd beweeggedrag en gewichtsbeheersing.

Bewijsmethode

SMS-score	5
Methode	Cluster-RCT
Sterke punten	• Geen beïnvloeding van de controlegroep door de interventiegroep vanwege geclusterde randomisatie.
Zwakke punten	• Vanwege personeelwisselingen (nieuwe instroom of switches tussen interventie- en controlevestigingen) kan er geen volledige vergelijking van baselinekarakteristieken gemaakt worden.

Toepasbaarheid

Context: De studie lijkt goed generaliseerbaar voor zover het directe effecten op gezond gedrag betreft (voor andere hoge-inkomenslanden), voor effecten op participatie en zorgkosten is de specifieke context in de VS van belang.

Werkzame elementen: Studie bevat een korte beleidstheorie: Aanbieden van workplace-wellness-activiteiten aan werknemers → Gezondere werknemers, lagere zorgkosten, minder verzuim en betere prestaties. De studie heeft daarnaast de intermediaire gedragsmechanismen via welke de interventie tot effecten op zorgkosten en gezondheid zou moeten leiden meegenomen. Meer inzicht in het proces zou eventueel bij kunnen dragen aan de toepasbaarheid.

Zijn er andere studies? Ja, maar deze tonen gemengde resultaten, soms effect op intermediair gedrag, maar weinig klinische of economische impact. Deze studie laat goed zien dat de resultaten van eerdere studies die veel meer positieve effecten vonden waarschijnlijk worden gedreven door selectie (relatief gezonde mensen participeren in het programma).

19. Head Start

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1), 159-208.

PICOT

P	Kinderen in arme counties in de VS in de jaren 60 ten tijde van de uitrol van Head Start
I	Hulp bij het aanvragen van Head Start. Head Start bestaat uit verschillende elementen, zoals voorschoolse opleiding, voeding en gezondheidszorg.
C	Geen hulp bij aanvragen van Head Start.
O	• Kindersterfte • Onderwijsuitkomsten
T	Uitkomsten worden op verschillende momenten gemeten tot de onderzochte kinderen volwassen zijn. De studie kijkt naar einduitkomsten.

Resultaten

In counties met een grotere kans op Head Start-financiering daalde de kindersterfte significant, terwijl tegelijkertijd de onderwijsuitkomsten verbeterden.

Bewijsmethode

SMS-score	4
Methode	Regression discontinuity design
Sterke punten	• Geen andere discontinuïteiten op de drempelwaarde.
Zwakke punten	• Enige verschillen in observeerbare karakteristieken tussen de interventie- en behandelgroep.

Toepasbaarheid

Context: Het artikel is zelf zeer terughoudend in de vertaling van de effectiviteit naar de huidige context in de VS. Belangrijkste reden is verschil in uitgangspositie (die nu voor deze doelgroep beter is dan tijdens het experiment in de jaren 60). Dit zal dus eveneens of zelfs meer gelden voor Nederland. Daarnaast is de vertaling naar een andere context uitdagend, omdat de studie in feite het effect van hulp bij een Head Start-aanvraag schat en niet van het meedoen aan het programma zelf.

Werkzame elementen: De studie bevat een korte beleidstheorie: Hulp bij aanvragen van Head Start → Meer gebruik van Head Start → Een verbetering van scholing, voeding en gezondheidszorg op jonge leeftijd, met name bij achtergestelde groepen met bijvoorbeeld weinig toegang tot zorg → Betere langetermijnuitkomsten via een accumulatie van levensloopeffecten (bv. betere opleiding). De studie brengt niet in kaart welke elementen van Head Start verantwoordelijk zijn voor de gevonden effecten op sterfte.

Zijn er andere studies? Ja, deze bevestigen vaak verbeterde cognitieve en gezondheidsuitkomsten, al variëren de langetermijneffecten.

20. WHO STOPS

Allender, S., Orellana, L., Crooks, N., Bolton, K.A., Fraser, P., et al. (2021). Four-Year Behavioral, Health-Related Quality of Life, and BMI Outcomes from a Cluster Randomized Whole of Systems Trial of Prevention Strategies for Childhood Obesity. *Obesity*, 29, 1022-1035. <https://doi.org/10.1002/oby.23130>

PICOT

P	Kinderen in groep 2, 4 en 6 (7-12 jaar) van basisscholen in 10 gemeenschappen in Zuidwest-Victoria, Australië
I	Whole-of-Systems (WHO) aanpak voor kindero obesitas met een community based-interventie. Het is een programma dat is ontwikkeld door relevante actoren in de betreffende gemeenschap en uit veel verschillende activiteiten bestaat, van regels aan ouders om alleen gezonde dranken mee te geven aan hun kinderen op school tot de aanleg van fietspaden om fysieke activiteit te stimuleren.
C	Gemeenschappen die later startten met de interventie.
O	• BMI • Prevalentie van overgewicht/obesitas • Leefstijlgedrag (afhaalmaaltijden, waterconsumptie) • Algemene gezondheid
T	Uitkomsten worden op verschillende momenten gemeten tot 4 jaar na start van de interventie. De studie meet zowel intermediaire uitkomsten als einduitkomsten.

Resultaten

Tussen 2015 en 2017 was er een significante daling in BMI en overgewicht, gevolgd door een stijging in de periode 2017-2019. Daarnaast verbeterde de gezondheidsgelateerde kwaliteit van leven (HRQoL) significant, zowel in het fysieke als in het psychosociale en globale domein. Op gedragsniveau kwamen meer gemengde effecten naar voren: zo nam de consumptie van afhaalmaaltijden af en steeg bij meisjes de waterinname, maar was er bij jongens juist een toename in het gebruik van energierijke snacks.

Bewijsmethode

SMS-score	5
Methode	Cluster-RCT (stepped wedge)
Sterke punten	• Geen beïnvloeding van de controlegroep door de interventiegroep als gevolg van geclusterde randomisatie.
Zwakke punten	• De invoering van de interventie was vertraagd voor de late starters. Het is onduidelijk of dit de resultaten heeft beïnvloed. • Uitkomsten zijn op clusterniveau onderzocht in plaats van op het niveau van het kind. • Karakteristieken van interventie- en controlegroep worden niet vergeleken.

Toepasbaarheid

Context: De uitkomsten zijn relevant voor vergelijkbare hoge-inkomenslanden. Stedelijkheid en de invloed van lokale instituties zijn punten van aandacht bij overzetten naar andere context.

Werkzame elementen: De studie bevat een korte beleidstheorie: Verschillende stakeholders uit de lokale gemeenschap maken zelf een probleemanalyse en ontwikkelen interventies → Gezamenlijk en integraal aanpakken van de verschillende elementen in de leefgemeenschap → Verbetering leefstijl en overgewicht bij kinderen. Er is in een ander paper een procesevaluatie gedaan voor een van de gemeenschappen (Allender et al., 2020). De werkzame elementen van de interventie worden niet direct in kaart gebracht.

Zijn er andere studies? Ja, maar er zijn weinig goede studies en het is onduidelijk wat verschillen tussen studies drijft.

