

Toekomstverkenningen - methodologie

In de applicatie NKR Cijfers zijn toekomstverkenningen van kankerincidentie en sterfte door kanker beschikbaar voor een periode van 20 jaar. Deze voorspellingen geven de modelmatig verwachte toekomstige incidentie en sterfte weer, ervan uitgaande dat de trend in de voorafgaande geobserveerde periode doorzet in de toekomst. Meer informatie over de toekomstverkenningen is te vinden op [naam van de landingspagina, met link naar de landingspagina].

Databronnen

Voor de toekomstverkenningen zijn gegevens uit de Nederlandse Kankerregistratie (NKR) en populatie- en sterftedata van het CBS gebruikt. We gebruiken NKR-data van mensen die sinds 1989 de diagnose kanker hebben gekregen. Van het CBS zijn de geobserveerde sterftedata en de geobserveerde populatiedata en voorspelde populatiedata voor de periode 2026 tot en met 2045 gebruikt. Meer informatie over de databronnen staat op de pagina [Uitleg NKR Cijfers](#) in de hoofdstukken incidentie en sterfte op deze webpagina.

Tumorsoortindeling en gemaakte keuzes

Voor de toekomstverkenningen van de kankerincidentie (nieuwe diagnoses) en sterfte door kanker gebruiken we dezelfde selectiecriteria en indeling van kankersoorten als voor de historische cijfers (cijfers tot nu) in NKR Cijfers. Meer informatie hierover is te vinden in het hoofdstuk over incidentie op de pagina [Uitleg NKR Cijfers](#).

Met de toekomstverkenningen laten we cijfers zien voor verschillende groepen van tumoren. De trend van een groep kan afwijken van de trends van de afzonderlijke tumoren binnen die groep. Voor sommige tumorsoorten of subgroepen hebben we geen toekomstverkenning gemaakt. Zo zijn er bijvoorbeeld geen toekomstverkenningen voor restgroepen: dit zijn gemengde groepen en met vaak weinig diagnoses. Ook ontbreken voorspellingen voor sommige andere subgroepen (zoals extreem zeldzame vormen van kanker), omdat er te weinig diagnoses zijn om de trend betrouwbaar te berekenen.

Voor de meeste tumorsoorten nemen we de gehele observatieperiode vanaf 1989 mee in de analyse. Maar voor sommige kiezen we ervoor om een kortere periode te gebruiken, bijvoorbeeld als er een duidelijke trendbreuk is. Dit kan komen doordat de registratie pas later is gestart (zoals bij basaalcelcarcinoom van de huid) of door veranderingen in de registratie (bijvoorbeeld bij myeloproliferatieve maligniteiten).

De filters voor de toekomstverkenningen in de applicatie wijken enigszins af van de filters voor de historische data. De toekomstverkenningen zijn beschikbaar voor 5 verschillende leeftijdsgroepen en per geslacht. We hebben geen toekomstverkenningen gemaakt naar stadium of provincie.

Statistische methoden

In onze toekomstverkenningen maken we gebruik van de geobserveerde incidentie en sterfte sinds 1989 tot op heden. Hiermee doen we voorspellingen van de verwachte incidentie en sterfte voor de komende twintig jaar.

In onze toekomstverkenningen doen we de fundamentele aanname dat de trend uit het verleden voorspellend is voor de toekomst. Onze methode maakt een voorspelling van toekomstige leeftijds- en geslachtsspecifieke incidentiecijfers per kalenderjaar. Deze vermenigvuldigen we vervolgens met de verwachte toekomstige omvang van de overeenkomstige leeftijds- en geslachtsspecifieke (midjaars-)bevolkingscijfers per jaar van het Centraal Bureau voor de Statistiek (CBS). Dit resulteert in de verwachte incidentie per geslacht, per leeftijdsgroep en kalenderjaar. Voor sterfte volgen we dezelfde procedure.

Modellen

Om leeftijd en geslachtsspecifieke incidentie en sterfte te voorspellen passen we het volgende algoritme toe:

- 1 We berekenen het leeftijds- en geslachtsspecifieke incidentiecijfer ('incidence rate') per kalenderjaar in de historische tijdreeks (voor de meeste kankers is dit vanaf 1989) in leeftijdsgroepen van 5 jaar. Dit is het aantal nieuwe kankerdiagnoses in een leeftijdscategorie gedeeld door de bevolking in die leeftijdscategorie.
- 2 We kijken in deze tijdreeks eerst of in de recente tijdreeks (de laatste tien geobserveerde kalenderjaren) de incidentie in alle jaren 0 was. In dat geval nemen we aan dat dit ook in de toekomst zo blijft.

- 3 In de overgebleven tijdreeksen kijken we of het incidentiecijfer laag was over de tijd, zo ja, dan schatten we het gemiddelde incidentiecijfer over de laatste 10 geobserveerde jaren en trekken we dit door naar de toekomst zonder trend. Het incidentiecijfer in een specifieke geslacht en leeftijdsgroep categoriseren we als 'laag' als 50% of meer van de jaarlijkse gemiddelde incidentiecijfers onder de 0.5 gevallen per 100 000 inwoners ligt.
- 4 Als het incidentiecijfer niet laag is, dan passen we *Bayesiaanse Gaussian Exponential Smoothing*-modellen toe om de trend te schatten en naar de toekomst door te trekken.

Exponential Smoothing-modellen

Een *Exponential Smoothing*-model met trend schat toekomstige datapunten op basis van het niveau van geobserveerde datapunten en de trend in die geobserveerde datapunten. Daarbij wegen het niveau van recente datapunten en de trend tussen recente datapunten zwaarder mee dan datapunten uit het verdere verleden. Het exacte gewicht dat aan recente datapunten versus oudere datapunten wordt toegekend, schat het model zelf. Er worden gewichten gebruikt die leiden tot de best mogelijke voorspelling van de waargenomen data. Het model bekijkt elk datapunt in de tijdreeks en probeert deze te voorspellen op basis van een gewogen combinatie van eerdere data, inclusief de trend in die eerdere data. Over het algemeen kunnen we stellen als niveau en trend stabiel zijn over de tijd, dat de historische tijdreeks veel gewicht krijgt. Daarentegen krijgen recente datapunten relatief meer gewicht als de kankerincidentie in de geobserveerde data sterk verandert op een onvoorspelbare manier.

Parameters

Bayesiaanse *exponential smoothing* modellen zijn als volgt geparametriseerd

$$\begin{aligned}
 y_{t,a} &\sim N(\hat{y}_{t,a}, \sigma_a) \\
 \hat{y}_{t,a} &= l_{t-1,a} + b_{t-1,a} \cdot \varphi_a \\
 l_{t,a} &= \alpha_a \cdot y_{t,a} + (1 - \alpha_a) \cdot (l_{t-1,a} + b_{t-1,a} \cdot \varphi_a) \\
 b_{t,a} &= \beta_a \cdot (l_{t,a} - l_{t-1,a}) + (1 - \beta_a) \cdot b_{t-1,a} \cdot \varphi_a
 \end{aligned}$$

- $y_{t,a}$ verwijst naar het leeftijdsspecifieke incidentiecijfer in jaar t en leeftijdsgroep a , en $N(\hat{y}_{t,a}, \sigma_a)$ naar een normaalverdeling met het verwachte incidentiecijfer $\hat{y}_{t,a}$ en standaardafwijking σ_a . In situaties waarin een dalende trend tot voorspellingen onder nul zou kunnen leiden, begrenzen we binnen het model voorspellingen onder nul, zodat de voorspelde incidentiecijfers niet onder nul kunnen gaan.
- De component $l_{t-1,a}$ wordt de 'niveaucomponent' genoemd en interpreteren we als de beste schatting van de verwachte waarde van het incidentiecijfer op tijdstip t voor leeftijdsgroep a , indien er geen trend zou zijn. De component $b_{t,a}$ wordt de trendcomponent genoemd en interpreteren we als de beste schatting van de verandering in incidentiecijfer die optreedt op tijdstip t voor leeftijdsgroep a .
- De parameters α_a en β_a zijn wegingsfactoren; hun waarden zijn begrensd tussen 0 en 1. Om een voorspelling te maken van het niveau $l_{t,a}$, gebruikt het model een gewogen gemiddelde van de geobserveerde waarde $y_{t,a}$ en de voorspelling $l_{t-1,a} + b_{t-1,a}$. Als de voorspelling minder goed overeenkomt met de geobserveerde gegevens, kent het model meer gewicht α_a toe aan de geobserveerde gegevens $y_{t,a}$. Wanneer daarentegen de observaties $y_{t,a}$ minder voorspellend zijn dan de modelvoorspelling, geeft het model meer gewicht $(1 - \alpha_a)$ aan de voorspelling $l_{t-1,a} + b_{t-1,a}$. Dezelfde logica geldt voor de schatting van de trend b met behulp van de wegingsfactor β_a .
- Omdat de schatting van de componenten $l_{t,a}$ en $b_{t,a}$ eerdere schattingen binnen de tijdreeks van l_a en b_a omvat, is elke schatting van $l_{t,a}$ en $b_{t,a}$ een gewogen gemiddelde van eerdere schattingen van deze componenten. Wanneer α_a en β_a dichter bij 0 liggen, kent het model meer gewicht toe aan eerdere schattingen; wanneer α_a en β_a dichter bij 1 liggen, krijgt het verleden minder gewicht bij het voorspellen van de toekomst.
- Ten slotte is de parameter φ_a een dempingsparameter; deze verlaagt het gewicht van de trendcomponent om een te sterk stijgende trend (wanneer de trendcomponent positief is) of een te sterke dalende trend (wanneer de trendcomponent negatief is) te voorkomen. Deze dempingsparameter schat het model ook zelf en is voor de meeste kankersoorten toegepast.

Zodra het model is geschat, kunnen we een voorspelling maken voor een toekomstig nog niet geobserveerd tijdstip t door trekkingen te nemen uit de *posteriori* verdeling van de parameters, evenals trekkingen uit een normaalverdeling met gemiddelde $\hat{y}_{t,a}$ en standaardafwijking σ_a om voorspelde waarden te bepalen. Deze waarden worden vervolgens iteratief gebruikt om de trend- en niveaucomponenten op $t+1$ te bepalen, enzovoort, tot het einde van de voorspelde tijdreeks.

Alle schattingen die uit het model voortkomen hebben een verdeling, de zogenaamde *posteriori*-verdeling, die de mate van (on)zekerheid weergeeft, conditioneel op de aannames die het model maakt. De gerapporteerde

puntschattingen corresponderen met de mediaan van de respectieve *posteriori*-verdelingen. Deze puntschattingen van de jaarlijkse leeftijds- en geslachtsspecifieke incidentiecijfers vermenigvuldigen we vervolgens met de verwachte omvang van de overeenkomstige leeftijds- en geslacht specifieke bevolkingscijfers in Nederland, gebaseerd op bevolkingsprognoses van het Centraal Bureau voor de Statistiek (CBS). Voor sterfte volgen we dezelfde procedure.

Betrouwbaarheid huidige en voorspelde cijfers

Bij voorspellingen komt per definitie onzekerheid kijken, zeker bij voorspellingen op lange termijn. Het is belangrijk om bij de interpretatie van de cijfers de beperkingen van de data en van de gebruikte analyses mee te nemen.

- De gegevens uit de Nederlandse Kankerregistratie NKR vormen de basis voor de analyses. Deze registratie is landelijk dekkend en nagenoeg compleet tot en met een jaar geleden. Een zeer kleine groep patiënten wordt niet geregistreerd, bijvoorbeeld wanneer er geen pathologische bevestiging en (poli)klinische diagnose is. Dit heeft geen invloed op de toekomstverkenningen.
- De sterftecijfers van het CBS kunnen afwijken van het werkelijke aantal overlijdens door verschillen in registratie en een zekere mate van subjectiviteit.
- De voorspelde aantallen nieuwe diagnoses projecteren we op de bevolkingsprognose van het CBS. Als de bevolking in werkelijkheid sneller of langzamer groeit dan verwacht, zullen de werkelijke aantallen afwijken van de voorspellingen.
- Het meest recente jaar in dat getoond wordt in NKR Cijfers is nog niet volledig. Daarom nemen we dit jaar niet mee in de analyses voor de toekomstverkenningen. De invloed hiervan is beperkt, vooral bij veelvoorkomende kankersoorten.
- In de toekomstverkenningen nemen we aan dat trends uit het verleden voorspellend zijn voor de toekomst. Dit zal niet altijd het geval zijn. Trends kunnen namelijk veranderen, door bijvoorbeeld ontwikkelingen in risicofactoren, screeningsprogramma's, diagnostiek, behandeling, preventie of door veranderingen in de registratie.
- Voor de toekomstverkenningen gebruiken we statistische modellen die veelgebruikt, internationaal erkend en geëvalueerd zijn¹. In de verkenningfase hebben we verschillende typen modellen met elkaar vergeleken onder uiteenlopende scenario's. Daarbij kwamen *Bayesian Gaussian Exponential Smoothing*-modellen als beste naar voren. Zoals voor alle modellen geldt, hebben ook deze methoden beperkingen, waardoor de voorspelde cijfers kunnen afwijken van de toekomstige werkelijkheid.

Bron

1. Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*, 3rd edition. OTexts, Melbourne, Australia. ISBN: 978-0-9875071-3-6.