

# Definering af buskundepotentiale ved spatiel regression af socioøkonomiske data og passagerforhold på Fyn

Thorbjørn Revsbech Sørensen<sup>1</sup>, Diana Andreea Vasile<sup>2</sup>, Patrycja Anna Zieba<sup>3</sup>, Flemming Albæk<sup>4</sup> og Seher Øzden<sup>5</sup>

<sup>1</sup>FynBus, Plan, ths@fynbus.dk, <sup>2</sup>FynBus, Plan, dva@fynbus.dk, <sup>3</sup>FynBus, Marked, paz@fynbus.dk, <sup>4</sup>FynBus, Økonomi og Analyse, fal@fynbus.dk og <sup>5</sup>FynBus, Økonomi og Analyse, soz@fynbus.dk

## Baggrund

Siden 2015 har FynBus benyttet en model der beskriver potentielle kunder på Fyn. Modellen er sammensat af socioøkonomiske data, og adfærdsdata som er definerede af forskellige registre og omdannet til et datasæt som hedder ConZoom™. Conzoom er et segmenteringsværktøj udviklet af Geomatic som klassificerer befolkningen i Conzoom-grupper og Conzoom-typer, så de hver især udgør så ensartede dele af den danske befolkning. Modellen kaldes i FynBus regi for Buskundepotentiale (BKP).

I 2023 valgte Geomatic at opdatere grundlaget for ConZoom™ og har i den forbindelse opdateret og tilføjet grupper og typer, som ikke eksisterer i den nuværende model. Der er derfor behov for en revidering af buskundepotentiale-modellen, især med henblik på de seneste opdateringer fra generation 5 til 6, som indeholder væsentlige ændringer, samt revurdering af passagerdata på baggrund af eksterne påvirkninger som f.eks. Corona-krisen. Dette medfølger at FynBus skal redefinere buskundeværdi (BKV) fra modellen fremadrettet med afsæt i generation 6 af ConZoom™ og FynBus' passagerdata post Covid-19.

BKP kortlægger potentielle buskunder med udgangspunkt i ConZoom™, demografiske data og faglige vurderinger. BKP definerer det samlede potentiale af buskunder, såfremt de ville have adgang til den bedst mulige rejse til deres individuelle behov. Det er altså ikke en model som estimerer et passagerantal, men estimerer potentiale ud fra bedste forudsætninger for individets rejse.

## Databeskrivelse

Conzoom data<sup>1</sup>, skabes af Geomatic og er et statistisk overblik over persontyper og grupper, som kan benyttes ved analyse, der involverer adfærd, målgruppearbejde, analyser som involverer socioøkonomiske elementer osv. Dataene skabes ud fra registre og sammensættes til typer og grupper. Der er i alt 9 grupper og 39 typer fordelt på disse grupper, som giver indsigt i adfærd i et spatielt afgrænset område. Conzoom datasættet repræsenterer, hvordan den danske befolkning så ud pr. 01-01-2023 og placerer husstandene i 100x100 kvadratmeters celler.

Hver type og gruppe er associeret med mange variabler, som er aggregeret på decilniveau eller er af typen boolean(sandt/falsk/dummy). Disse dækker eks. over boligtype, civilstand, bilejerskab, familiestørrelse, økonomi, fritidsbeskæftigelser, uddannelsesniveau m.m. Der er indlagt et eksempel på en gruppe og en type i Bilag 1. ellers kan oversigtens findes på ConZooms hjemmeside.

---

<sup>1</sup> <https://www.conzoom.dk/da/danmark/segmenter>



indenfor cellernes afgrænsninger. For stoppesteder uden påstigere registreret via Rejsekort, er Conzoom-data koblet med PTS-data (Passager Tælle System), som er baseret på realtidstællinger foretaget af kameraer i busserne. Dette er for at dække evt. mangler som følge af rejsekortprodukter som ikke kræver check-ind/ud og giver et mere repræsentativt datasæt. Det anvendte påstigningsdata er på stoppested- og månedsniveau for 2023. Datasættet er ikke afgrænset til kun at indeholde busstoppesteder, men består dermed af data for alle stoppesteder, dvs. letbane- og togstationer er også inkluderet, såfremt der har været nogen registrerede påstigninger i 2023.

Formålet med dette datasæt er at anvende det i en regressionsanalyse for at generere nye BKV'er. Disse værdier indgår i beregningsformlen for BKP og bruges til at kvantificere hver enkelt Conzoom-gruppe behov for og/eller villighed til at benytte kollektiv trafik. Med andre ord estimerer BKV sandsynligheden for, at hver Conzoom-gruppe rejser med kollektiv transport. For at opnå de mest præcise værdier er alle relevante variabler, der forventes at påvirke brugen af kollektiv trafik, inkluderet i analysen.

## Metode

BKP defineres ud fra følgende formel:

$$\text{Buskundepotentiale} = \text{Fra Potentiale} + \text{Til Potentiale} \quad (1)$$

$$\text{Fra Potentiale} = (\text{Natbefolkning}_{\text{AntalPersoner}}) \cdot (\text{BusVærdi FynBus}) / 100 \quad (2)$$

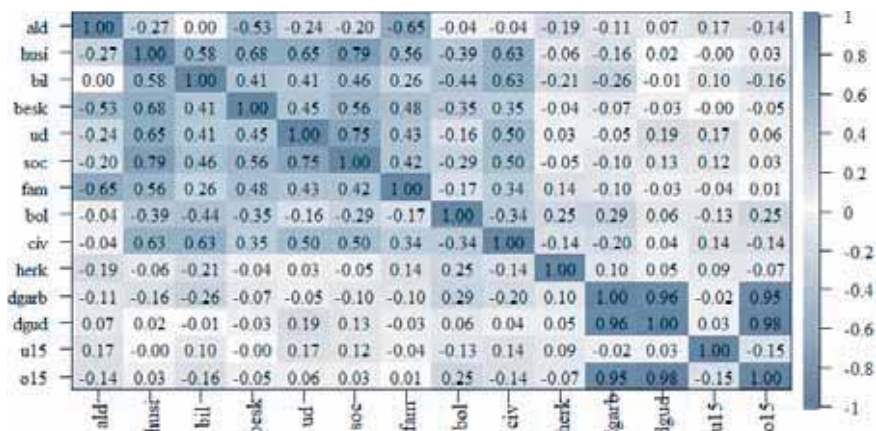
$$\text{Til Potentiale} = (\text{Dagbefolkning}_{\text{Beskaeftigede}} \cdot 0,2) + (\text{Dagbefolkning}_{\text{UddannelseOver15}} \cdot 0,52) \quad (3)$$

Natbefolkning og dagbefolkning er defineret af Geomatic på baggrund af statistiske registre, og busværdien er fastsat baseret på gruppen og typernes forbrugsadfærd, som er udarbejdet af Geomatic, Movia og FynBus.

FynBus har valgt at benytte spatiel regression til at forsøge at definere en ny BKV, da teknologien tillader muligheden for langt mere detaljerede passagertal i forhold til 2015. Dette skaber grundlaget for en bedre og mere nøjagtig model.

Arbejdet med ConZoom™ data og stoppested-/passagerdata kræver at der tages højde for den spatielle dimension. Passagertallene tager udgangspunkt i historiske data fra 2023, hvor der er benyttet data fra rejsekortet, passagertælleudstyr i busserne og data fra FynBus egne salgskanaler. Den tværfaglige gruppe, bestående af repræsentanter fra Markeds-, Planlægnings- og Data-/Analyseafdelingerne, har udvalgt ConZoom™-decilvariabler, der vurderes relevante for adfærd, som påvirker villigheden til at benytte kollektiv trafik.

De udvalgte decilvariabler blev analyseret i en indledende undersøgelse, hvor korrelationerne mellem variablerne blev afdækket til brug i den endelige model. Resultaterne blev samlet i en Spearman-korrelationsmatrice for at identificere og ekskludere variabler med høj indbyrdes korrelation. Der er ingen signifikante outliers, hvilket vil sige, at regressionsmodellerne er kørt på det komplette datasæt. Dog er datasættet afgrænset til kun at bestå af de celler, hvor der har været påstigere/pts-aktivitet, resten er sorteret fra. Resultatet er som følger:



Figur 1: Forkortelser i korrelationsmatricen. ald: aldersdecil, husi: husstandsindkomstdecil, bil: bilrædighedsdecil, besk: beskæftigelsesdecil, ud: uddannelsesdecil, soc: socialklassedecil, fam: familietypedecil, bol: boligtypefaktor, civ: civilstandsdecil, herk: herkomstfaktor, darb: dagbefolkning i arbejde, dgud: dagbefolkning i uddannelse, uu15: uddannelse under 15, uo15: uddannelse over 15.

Ud fra ovenstående matrice er decilvariablerne alder, husstandsindkomst, bilrædighed, beskæftigelse, uddannelse, socialklasse, familietype, boligtype, herkomst og civilstand anvendt i regressionsmodellen. Efterfølgende har regressionskoefficienterne for familietype, herkomst og boligtype vist sig ikke at være signifikante, hvorfor de er blevet taget ud i den endelige model.

Derfor er der undersøgt følgende tre modeller til at udføre spatiel regression på BKV, da disse modeller ofte ses benyttelse til lignede spatielle analyser.

Spatiel Lag Model (SLM), der også kendes som Spatial Autoregressiv Model er kendetegnet ved at den afhængige variable er påvirket af nærliggende celler.

$$y = \rho W y + X \beta + \varepsilon \quad (4)$$

hvor  $y$  er den afhængige variabel,  $\rho$  måler graden af spatiel afhængighed,  $W$  er den spatielle vægtmatrice,  $X$  er de uafhængige variable,  $\beta$  er koefficienterne og  $\varepsilon$  er fejleddet.

Spatiel Error Model (SEM) er kendetegnet ved at der tages højde for korrelationer i fejleddet, der kan være til stede på grund af udeladte variable.

$$y = X \beta + u, \quad \mu = \lambda W u + \varepsilon \quad (5)$$

hvor  $y$  er den afhængige variabel,  $X$  er den uafhængige variabel,  $\beta$  er koefficienterne,  $\lambda$  er den spatielle fejleddet, som repræsenterer graden af den spatielle korrelation i fejleddet,  $W$  er den spatielle vægtmatrice med forholdet mellem cellerne og  $u$  er en vægtmatrice af uafhængige fejleddet.

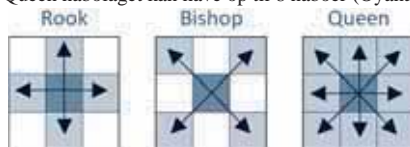
Spatiel Durbin Model (SDM) er en udvidet model af SLM, hvor både den afhængige og de uafhængige variable er påvirket af de nærliggende celler.

$$y = \rho W y + X \beta + W X \theta + \varepsilon \quad (6)$$

hvor  $y$  er den afhængige variabel,  $X$  er den uafhængige variabel,  $\beta$  er koefficienterne,  $\theta$  er koefficienterne for de spatielle lags af de uafhængige variable,  $W$  er den spatielle vægtmatrice med forholdet mellem cellerne og  $\varepsilon$  er fejleddet (Anselin, 1988).

For at udføre en spatiel regression skal der i modellen defineres afgrænsning i form af nabolag, der spatielle regression modeller som udgangspunkt er af tre typer af nabolag. Disse, baseret på skakbrikker, er kendt som "Rook", "Bishop" og "Queen".

- Rook: betragter kun naboer som er direkte horisontale eller vertikale ud fra cellens placering. Rook nabolaget kan have op til 4 naboer.
- Bishop: betragter kun naboer som ligger på de diagonale akser ud fra cellens placering. Bishop nabolaget kan have op til 4 naboer.
- Queen: betragter naboer som ligger på både diagonale, vertikale og diagonale akser ud fra cellens placering. Queen nabolaget kan have op til 8 naboer (Oyana, 2021).



Figur 2: Illustration af de tre forskellige nabo typer, der kan vælges mellem

Det vurderes at nabolaget Queen er bedst egnet til regression analysens formål, hvorfor der ikke laves nogen modeller med Rook eller Bishop-nabolaget. Dog er der i ConZoom™ observationer som ikke har nogen nabo, fx i kantområder eller isolerede områder. For at undgå at modellen ikke udelukker disse celler er der i modellen specificeret at disse celler skal medtages. Dette giver et mere omfattende billede af sammenhængen mellem påstigninger i bussen og socioøkonomiske faktorer i områder med få naboer. Det besluttes at regressionsmodellen SEM benyttes, hvor nabolaget defineres som queen, idet der vurderes det vil give de mest retvisende resultater, eftersom nabolaget queen inkluderer alle naboer omkring en celle.

Den endelige model defineres derfor som følgende:

$$\begin{aligned} \log(\text{BusPåstignere}) &= \beta_0 + \beta_1 \text{alder} + \beta_2 \text{civilstand} + \beta_3 \text{boligtype} + \beta_4 \text{husstandsindkomst} \\ &+ \beta_5 \text{bilrådighed} + \beta_6 \text{beskæftigelse} + \beta_7 \text{uddannelse} + \beta_8 \text{socialklasse} \\ &+ \text{Conzoom grupper} + \varepsilon \end{aligned} \quad (7)$$

Regressionskoefficienterne som skabes af denne model skal efterfølgende standardiseres til en værdi der kan indgå i beregningen for BKP. Formlen for BKV er designet til at rumme værdier som er på hele tal og på en skala fra 0 – 100. For at kunne indsætte vores nye værdier ind i den eksisterende formel, er det besluttet at standardisere resultaterne så de tilpasses denne skala. Resultaterne fra regressionsanalysen er derfor blevet omregnet til en vægt ved følgende formel:

$$Vægt = \frac{\text{koefficient} - \text{koefficient}_{\min}}{\text{koefficient}_{\max} - \text{koefficient}_{\min}} \Leftrightarrow \frac{\text{koefficient} - -0,85}{2,06 - -0,85} \quad (8)$$

Vægtene er efterfølgende ganget med 100 og afrundet til nærmeste hele tier, så de kan anvendes som BKV i formlen "Fra Potentiale". I nedenstående tabel fremgår de beregnede BKV. Buskundefærdien for Gruppe A. C og F er justeret på baggrund af tidligere model, og vurdering fra Markedsafdelingen.

## Resultater

Analysen af den spatielle regression på buskundefærdi har resulteret i koefficienter som er vægtet for at give os en heltalsværdi som der kan benyttes i den eksisterende formel for buskundefærdi. Denne

nye værdi erstatter den forrige, hvor grundlaget for værdien er ukendt, hvilket fremgår af nedenstående Tabel 1.

For at kunne sammenligne med den nye model som er dannet på baggrund af gruppeniveau med buskundefærdierne fra den eksisterende model som er dannet på typeniveau fremgår middelværdien for BKV for gruppen. Dette gøres for at vurderer om værdierne udskiller sig væsentligt fra den tidligere model.

Gruppe kode	Gruppe navn	Koefficient	Vægt	Afrundet BKV	Middel af BKV i dag
A	Velfunderede husejere	-0,38	0,16	20	8
B	Komfort og hygge	0,24	0,38	40	16
C	Lever på landet	-0,85	0,00	5	5
D	Velhavere	0,08	0,32	30	10
E	Urban mangfoldighed	0,38	0,42	40	58
F	Vid og velstand	2,06	1,00	90	36
G	Seniorer	0,44	0,44	40	16
H	Unge på vej	0,38	0,42	40	80
I	Provinsliv	0,53	0,47	50	35

Tabel 1: Resultater for Spatiel Error Regression på ConZoom™ gruppeniveau. Kolonnen "Middel af BKV i dag" er den middelværdien af buskundefærdierne på gruppeniveau.

Da grundlaget er skiftet fra type til gruppe, er der selvfølgelig færre nuancer i buskundefærdien end der var tidligere, hvilket middelværdien også giver udtryk for. Dog ved test af dataene er Buskundefærdierne på det spatielle niveau tilsvarende til den eksisterende model. Derfor er det vurderet at resultaterne er brugbare i det videre arbejde med buskundefærdier hos FynBus.

Der vurderes undersøgt om rejsehjemfordelinger skal benyttes til at bestemme faktorerne der anvendes i formen Til-Potentiale (3). Der er dog opstillet fire scenarier der kan visualisere hvordan disse faktorer påvirker det endelige buskundefærdier. I disse scenarier indgår en følsomhedsanalyse hvor vi har inddelt faktorerne. Et hvor de oprindelige faktorer er nedskaleret med henholdsvis 5 og 10 procentpoint. Samt et scenarie som kigger på andelen af beskæftigede, og folk under uddannelse, ud af den samlede befolkning.

Scenarierne tager derfor udgangspunkt i den oprindelige models afgrænsninger hvor det er antaget at 20% af alle i beskæftigelse er potentielle kunder og 52% af uddannelsessøgende over 15 er potentielle kunder

**Scenarie 1:** tager udgangspunkt i modellens tidligere værdier, hvor 20% af hele dagbefolkningen er potentielle kunder og 52% af alle studiesøgende er potentielle kunder.

**Scenarie 2:** tager udgangspunkt i scenarie 1, men nedskriver begge værdier med 5 procentpoint.

**Scenarie 3:** tager udgangspunkt i scenarie 1, men nedskriver begge værdier med 10 procentpoint.

**Scenarie 4:** tager udgangspunkt i en optimistisk skalering som tager udgangspunkt i det totale potentiale af hele Fyns befolkning, det er antaget denne model at alle beskæftigede og alle under uddannelse er potentielle kunder. Dette indeholder værdierne 43% for dagbefolkningen og 12% af befolkning i uddannelse over 15.

Nedenstående Tabel 2 viser umiddelbare resultater fra følsomhedsanalysen.

Scenarie	Til potentiale	Dagbefolkning i beskæftigelse	Dagbefolkning i uddannelse over 15 år	BKP	Ændring i %
1	20% og 52%	65.494	31.473	231.715	38%
2	15% og 47%	52.619	28.447	215.814	28%
3	10% og 42%	39.744	25.421	199.913	19%
4	43% og 12%	97.235	7.360	239.343	42%

Tabel 2: I ovenstående tabel fremgår resultaterne fra følsomhedsanalysen. BKP står for buskundepotentiale.

Scenarie 1, 2 og 3 vurderes at kunne arbejdes videre med som grundlag for den opdaterede buskundemodell, alt efter interne målsætninger hos FynBus.

## Diskussion og konklusion

Spatiel regressionsanalyse som metode til at definere buskundeværdien til brug i formen for buskundepotentiale, viser lovende resultater til videre udvikling. Differencen mellem de nye buskundeværdier og buskundeværdierne defineret i den nuværende model for buskundepotentiale viser sig for nogen grupper større end forventet. Derfor er det påtænkt at modellen bør testes på typer fremfor grupper ved fremtidig udvikling inden modellen fuldt udrulles.

Resultaterne er som forventet højere end den nuværende model, netop grundet de mange variationer som der fordeler sig på det mere detaljerede typeniveau for ConZoom™ data. Typeniveauet er dog meget ressourcekrævende at afdække, og har ikke været muligt under projektets udformning. Væksten i buskundepotentialet er dog i tråd med stigende tilflytning som opleves på Fyn gennem de sidste 10 år, hvori det er forventet at der tilsvarende er et stigende potentiale for buskunder.

Herudover, ønskes det afdækket hvorvidt faktorerne i Til-Potentiale (3) skal vurderes ud fra andre metoder, herunder udnyttelse Transportvaneundersøgelsen, som udarbejdes af DTU, eller andre statistiske modeller der kortlægger rejser på baggrund af transportmiddelfordeling.

Desuden, bør der vurderes hvorvidt decilindelingen hos ConZoom™ dataene er for brede i forhold til formålet, her kan det overvejes at supplere med yderligere data fra Geomatic, eller Danmarks statistik som indeholder variabler med kontinuerlige variabler eller andre tilsvarende variabler. Begge disse muligheder kan potentielt forbedre modellen, dog er udnyttelse af realtidsdata for passagertællinger med kameraer i busserne vurderet som meget brugbare som en afhængig variabel for denne type regression.

## Konklusion

Modellen, der anvender spatial regressionsanalyse, viser lovende resultater. Det er dog nødvendigt at undersøge modellens resultater på typeniveau for at sikre, at nuancerne i data er tilstrækkeligt belyst, og at værdierne ikke overvurderes i forhold til den eksisterende model.

Derfor fortsættes arbejdet med fokus på at afdække disse usikkerheder, inden modellen implementeres som et analyseværktøj hos FynBus.



## Litteratur

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academics, Dordrecht

Oyana, T. J. 2021. *Spatial analysis with R: Statistics, visualization, and computational methods* (2nd ed). CRC Press. <https://doi.org/10.1201/9781003021643>

## Bilag 1: ConZoom™ data

De anvendte variabler er decilvariabler, der inddeler husstande på Fyn i ti lige store grupper<sup>2</sup>. Disse variabler som er brugt i analysen, ser ud som følgende:

Decil	Alder	Husstandsindkomst	Bilrådighed	Beskæftigelse	Uddannelse	Socialklasse	Familietype	Boligtype	Civilstand
1	0-29 årige	Lav	Primært husstande uden bil	Primært pensionerede	Primært husstande med grundskole	Meget stor andel af arbejderklasse og underklasse	Primært enlige uden børn	Primært enfamiliehus og stuehus	Primært enlige
2	30-49 årige		Primært husstande uden bil	Primært udenfor arbejdsmarkedet	Primært husstande med erhvervsfaglig og grundskole	Stor andel af arbejderklasse og underklasse	Primært enlige uden børn	Primært enfamiliehus	Primært enlige
3	30-49 årige		Primært husstande uden bil	Primært Lønmodtagere på lavt eller mellem niveau	Primært husstande med erhvervsfaglig og grundskole	Primært middelklasse og arbejderklasse	Primært enlige uden børn	Primært enfamiliehus	Primært enlige
4	30-49 årige		Primært husstande med mindst en bil	Primært Lønmodtagere på lavt eller mellem niveau	Primært husstande med erhvervsfaglig uddannelse	Primært middelklasse og arbejderklasse	Primært enlige uden børn	Primært enfamiliehus	Primært enlige
5	49-59 årige		Primært husstande med mindst en bil	Primært Lønmodtagere på lavt eller mellem niveau	Primært husstande med erhvervsfaglig uddannelse	Stor andel af middelklasse og højere middelklasse	Primært enlige uden børn	Primært enfamiliehus og række/kæde hus	Primært enlige
6	Primært personer over 30 år		Primært husstande med mindst en bil	Primært Selvstændig eller Lønmodtagere på højt niveau	Primært husstande med erhvervsfaglig uddannelse	Stor andel af middelklasse og højere middelklasse	Primært med børn	Primært række/kæde hus	Primært par
7	Primært personer over 30 år		Primært husstande med mindst en bil	Primært Selvstændig eller Lønmodtagere på højt niveau	Primært husstande med erhvervsfaglig uddannelse	Primært middelklasse eller over	Primært med børn	Primært etagebolig	Primært par
8	Primært personer over 65 år		Primært husstande med mindst en bil	Primært Selvstændig eller Lønmodtagere på højt niveau	Primært husstande med erhvervsfaglig uddannelse	Primært middelklasse eller over	Primært med børn	Primært etagebolig	Primært par
9	Primært personer over 65 år		Primært husstande med mindst en bil	Primært Selvstændig eller Lønmodtagere på højt niveau	Primært husstande med mellem og lang videregående uddannelse	Stor andel af overklasse og højere middelklasse	Primært par med børn	Primært etagebolig	Primært par
10	Primært personer over 65 år	Høj	Primært husstande med mindst en bil	Primært Selvstændig eller Lønmodtagere på højt niveau	Primært husstande med lang videregående uddannelse og forskere	Primært overklasse	Primært par med børn	Primært etagebolig	Kun par

<sup>2</sup> <https://apps.conzoom.eu/api/v1/meta/dk/variables>



## **Aid for Trade and CO2 Emissions: The Case of Middle-Income Countries**

**Ayşe Arı, Department of Economics, Mersin University, Jørgen T. Lauridsen, Department of Economics, University of Southern Denmark, and Elvan Küpeli, Department of Economics, Mersin University. Correspondence: [jtl@sam.sdu.dk](mailto:jtl@sam.sdu.dk).**

### **Abstract**

We investigate the impact of Aid for Trade (AfT) on the environment for 12 middle-income countries during the period of 2003-2020 by employing the Augmented Mean Group estimator and causality tests. While the direct effect of AfT on CO2 emissions is significant and negative for the renewable as well as the total energy sector, the indirect effect is not. We didn't find evidence of an inverted-U-shaped effect of AfT on CO2 emissions. A causality relationship between AfT and CO2 emissions was found for the total as well as the renewable energy sectors. Thus, AfT funds targeted to renewable or total energy sectors in middle-income countries can reduce CO2 emission but not AfT funds for other sectors.

**Keywords:** Sustainable development; Aid for Trade; CO2 Emissions; Trade; Renewable Energy; Economic Development; Causality

### **1.Introduction**

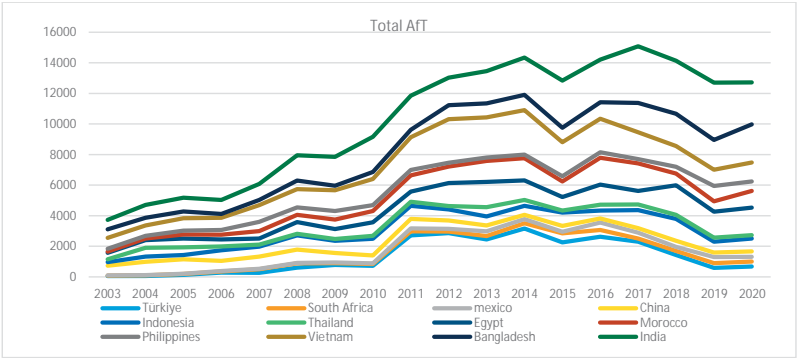
In the present study, we analyse the impact of AfT on the environment for middle income countries. Detecting an improvement in environmental quality resulting from AfT disbursements in these countries will indicate that in addition to the main purpose of AfT, which is increased trade benefits, AfT is also beneficial for environment. Another contribution of this study is to analyse the direct and indirect impact of AfT on CO2 emissions. To this aim, we separate AfT into two groups depending on sectoral basis: a) AfT funds transferred to the total energy sector to reflect the direct impact of AfT, and b) AfT funds transferred to sectors other than the energy sector in order to reveal the indirect effects of AfT. AfT inflows for sectors other than the energy sector are mostly provided to the production sectors and infrastructure investments such as communication, transportation and storage. In this study we search answers to the following questions: Do AfT funds for production and infrastructure investments have

any effect on environmental quality through macroeconomic variables such as economic growth and poverty and, if so, is this effect positive or negative, and do the AfT inflows for the total energy sector or the renewable energy sector have any impact on environmental quality? One finding revealing that AfT funds going to the energy sector increase the environmental quality indicates that these funds were used mostly in the renewable energy sector, not in the fossil fuel energy. As stated by Kablan and Chouard (2022:5358), foreign aid for renewable energy is provided to support projects for renewable energy resources and production. Therefore, AfT disbursements for the renewable energy sector will be expected to improve environmental quality more than other forms of AfT. In the present study, the validity of this thesis will be analysed. Lastly, we investigated the existence of an inverted U-shaped relationship between total AfT and environmental quality, as Boly (2018:18) states that the effect of foreign aid on the environment quality depends on the monitoring power of donor countries and the aim of authorities in recipient countries which is related to the magnitude of foreign aid. This may also be valid for the relationship between AfT and CO<sub>2</sub> emissions. In addition, AfT may have effect on the environment through different channels such as economic growth, poverty, income inequality etc. The effects of these macroeconomic variables may realize in the opposite direction which can also change after a threshold value of AfT. In the present study we aim to test the validity of U shaped or inverted U-shaped AfT hypothesis between AfT-CO<sub>2</sub> emissions.

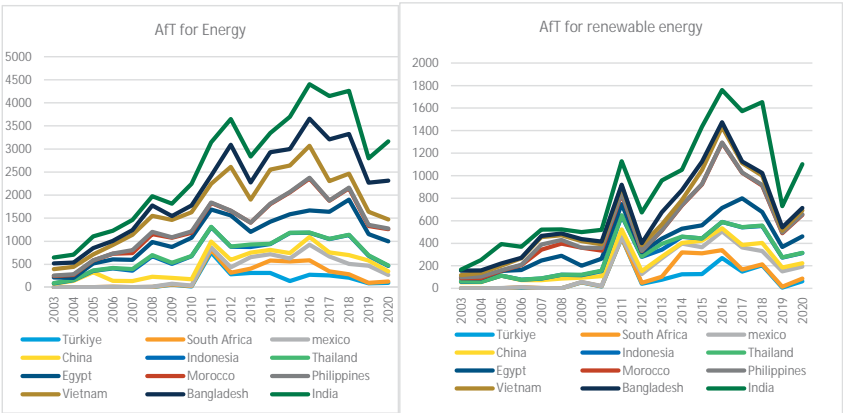
Data from 12 middle income countries (Turkey, South Africa, Mexico, India, Morocco, Bangladesh, Philippines, Vietnam, Egypt, Thailand, Indonesia, and China) are analysed covering the years from 2003 to 2020. In the graph below the total AFT inflows in US\$ are showed for the mentioned countries. According to the Figure 1, the highest AfT funds go to India, Bangladesh, Vietnam, Philippines, Morocco, and Egypt. The lowest AfT inflows belong to Türkiye, South Africa, Mexico and China. This makes sense since Türkiye, South Africa, Mexico and China are relatively more developed than the other countries in the study. It is also seen that the countries receiving most and least AfT funds have almost not changed during the years 2003 to 2020. In addition, we can say that total AfT has generally increased since 2003 and that the peak values were reached in the years 2014 and 2017. Figures 1 and 2 also illustrate that there has been an

important increase in AfT funds, AfT for total energy and AfT for renewable energy sectors during the years 2010 to 2019. Lastly, following a reduction in the years 2015 and 2019, these funds have started to increase again for most of the recipient countries in the model.

**Figure 1. Total AfT Funds**



**Figure 2. AfT disbursements for energy sector and renewable energy sector**



In Figure 2, AfT inflows to the total energy sector and renewable energy sector (US \$) are displayed. We see that India, Bangladesh, Vietnam, Philippines and Egypt have the highest AfT inflows to total energy sector. Türkiye, South Africa, Mexico and China

have the lowest AfT inflows for energy sector which is parallel with the results of total AfT.

When we look at the AfT funds for the renewable energy sector, Figure 2 shows that India, Bangladesh, Vietnam, Philippines and Egypt have the highest funds. It is also seen that the countries receiving the lowest AfT inflows for renewable energy sector are again Türkiye, South Africa, Mexico and China as in the case of total energy sector.

#### 4. Model and Data

Model 1, which aims to determine the direct and indirect influence of AfT on the environment, is written as (1) following the study of Nguyen et al. (2023):

$$\text{Model 1: } CO2_{it} = \beta_1 AfT_{ener,it} + \beta_2 AfT_{other1,it} + \beta_3 GDP_{it} + \beta_4 renew_{it} + e_{it} \quad (1)$$

In order to investigate the direct impact of AfT on the environment, we preferred AfT inflows for the renewable energy sector ( $AfT_{ren}$ ) as the second variable. For this purpose, we created Model 2. Similar to Model 1, the  $AfT_{other2}$  covers the AfT funds for all sectors except renewable energy sector:

$$\text{Model 2: } CO2_{it} = \beta_1 AfT_{ren,it} + \beta_2 AfT_{other2,it} + \beta_3 GDP_{it} + \beta_4 renew_{it} + e_{it} \quad (2)$$

Finally, in the study, the impact of total AfT on environment was also analysed for developing countries. In this Model 3, the possibility of the impact of AfT on the environment being U or inverted-U shaped was also considered:

$$\text{Model 3: } CO2_{it} = \beta_1 AfT_{it} + \beta_2 AfT_{it}^2 + \beta_3 GDP_{it} + \beta_4 renew_{it} + e_{it} \quad (3)$$

To determine the relationship between environmental quality and AfT, the study include 12 developing countries (Turkey, South Africa, Mexico, India, Morocco, Bangladesh, Philippines, Vietnam, Egypt, Thailand, Indonesia and China) that are in the middle-income group according to the World Bank classification. Data covers the period 2003-2020 in the analyses. While the data for AfT inflows were taken from the OECD site (OECD, 2024), other data were obtained from the World Bank (WDI, 2023). The data were used in logarithmic forms.

## 5. Empirical results

In this study, firstly, a cross-sectional dependency test was conducted. Determining cross-sectional dependency provides guidance in selecting estimators and cointegration and causality tests in order to produce consistent and reliable results. The LM test of Breusch and Pagan (1980) and the  $LM_{adj}$  test of Pesaran et al. (2008) were preferred which can be used in the case of  $T > N$ .  $LM_{adj}$  test is a stronger version of the LM test, and Pesaran et al. (2008) developed it by including the mean and variance to the LM test.

The null hypothesis indicates no cross-sectional dependence for  $LM_{adj}$  test. According to the test results summarized in Table 2, the probability values are less than 0.05, thus revealing cross-sectional dependency.

**Table 2. Cross-Sectional Dependency Test Results (p-values in parentheses)**

	LM	$LM_{adj}$
CO2	634.079 (<0.001)	14.816 (<0.001)
AfT <sub>ener</sub>	248.649 (<0.001)	9.292 (<0.001)
AfT <sub>other1</sub>	228.931 (<0.001)	14.566 (<0.001)
AfT <sub>ren</sub>	241.254 (<0.001)	5.552 (<0.001)
AfT <sub>other2</sub>	237.222 (<0.001)	13.999 (<0.001)
AfT	262.655 (<0.001)	13.973 (<0.001)
AfT <sup>2</sup>	262.4683 (<0.001)	13.598 (<0.001)
renew	480.5656 (<0.001)	18.039 (<0.001)
gdp	973.2650 (<0.001)	11.697 (<0.001)

Whether the series contained unit roots was investigated with the Cross-sectionally Augmented Dickey Fuller (CADF) test developed by Pesaran (2007). The CADF test is one of the second-generation unit root tests that considers cross-section dependency. Pesaran (2007) reached Cross-sectionally Augmented IPS (CIPS) statistics that give results for the entire panel by taking the average of the CADF test statistics obtained for each cross-section and adding the lags of the averages to the model. The CIPS test has powerful size properties for small samples and can produce consistent results both in the cases of  $N > T$  and  $T > N$  (Pesaran, 2007:266-267). Unit root test results are summarized in Table 3 for the intercept model.

Table 3 indicates that while the variables  $CO_2$ ,  $AfT_{other2}$ ,  $AfT$  and  $renew$  are difference stationary,  $AfT_{ener}$ ,  $AfT_{other1}$ ,  $AfT_{ren}$ ,  $AfT^2$  and  $GDP$  are level stationary. The results of unit root tests indicate that the dependent variable is I(1) while some of the independent variables are I(1) and the others are I(0).

**Table 3. Results of Unit Root Test**

	Level	First diff.
	Test stat.	Test stat.
CO2	-1.663	-3.161
$AfT_{ener}$	-2.901	-5.045
$AfT_{other1}$	-2.271	-4.329
$AfT_{ren}$	-2.719	-3.995
$AfT_{other2}$	-2.052	-4.613
$AfT$	-2.118	-4.475
$AfT^2$	-2.692	-4.448
renew	-1.573	-3.316
gdp	-2.431	-3.529

Note: CIPS critical values for 10%, 5% and 1% significance levels are -2.14, -2.26 and -2.47, respectively. Critical values are obtained from Pesaran (2007:280).

It is also useful to know whether our models are homogeneous or heterogeneous. Determining whether the slope coefficients in the cointegration equation are the same among the countries is also decisive in choosing appropriate cointegration tests, coefficient estimators and causality tests. For this purpose, the homogeneity test of Pesaran and Yamagata (2008) was applied. This test was conducted as a development of the Swamy (1970) test statistic ( $\tilde{S}$ ), and can be used for samples where  $N < T$  as well as samples with  $T > N$ . Two further tests from the study of Pesaran and Yamagata (2008) are applied. Among the tests, the  $\tilde{\Delta}$  test produces consistent results for large samples. The  $\tilde{\Delta}_{adj}$  test is the bias-adjusted version and produces more consistent results for small samples (Pesaran and Yamagata, 2008:57).

Homogeneity test results are presented in Table 4. According to Table 4, since the p-values are less than 0.05, homogeneity is rejected.

**Table 4. Pesaran and Yamagata (2008) homogeneity test results**

	Model 1		Model 2		Model 3	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
$\tilde{\Delta}$	11.273	0.000	11.290	0.000	11.170	0.000
$\tilde{\Delta}_{adj}$	13.807	0.000	13.827	0.000	13.681	0.000

In order to investigate the existence of a long-run relationship between variables we applied the Durbin Hausman technique of Westerlund (2008) which allows the explanatory variables to be I(0) or I(1). Additionally, this test is used for samples with cross-sectional dependence. Another advantage of the Durbin Hausman technique is that it can produce results for both homogenous and heterogenous panels. Related to this test, group statistics demonstrates results for heterogenous models while panel statistics demonstrates results for homogenous models. (Westerlund, 2008). Table 5 below gives the results of these tests for Model 1, Model 2 and Model 3. According to these results, the p-values are greater than 0.05 which denote that the series are not cointegrated for all 3 models.

**Table 5. Cointegration Test Result**

	Model 1		Model 2		Model 3	
	Stat.	p-value	Stat.	p-value	Stat.	p-value
dh_g	-0.222	0.412	1.498	0.933	0.786	0.784
dh_p	0.009	0.503	-0.007	0.497	-0.064	0.475

Although the variables are not cointegrated we can search for coefficient estimates of the regression models by employing the above-mentioned AMG method of Eberhardt and Bond (2009). Thus, the direction and magnitude of the direct and indirect impact of AfT on the environment can be determined. The AMG method allows cross-sectional dependency and heterogeneity. Another advantage of the AMG method is that it is demonstrated by Monte Carlo simulations that results are unbiased whether the variables are cointegrated or not (Eberhardt, 2012:64). Similar to the Common Correlated Effects (CCE) method, the AMG estimator will also be valid if there is unit root process in regressors (Pesaran, 2007:969-75; Eberhardt and Bond, 2009; Eberhardt and Teal,



2010). In short, the AMG estimator also allows the integration of explanatory variables at different degrees (Purcel, 2021:8-72).

**Table 6. Results of AMG estimator results**

	<b>Model 1</b>		<b>Model 2</b>		<b>Model 3</b>	
	<b>Coef.</b>	<b>p-value</b>	<b>Coef.</b>	<b>p-value</b>	<b>Coef.</b>	<b>p-value</b>
AfT <sub>ener</sub>	-.0157894	0.006				
AfT <sub>other1</sub>	-.0097958	0.448				
AfT <sub>ren</sub>			-.0062864	0.003		
AfT <sub>other2</sub>			-.0157805	0.282		
AfT					-.2106083	0.674
AfT <sup>2</sup>					-.0144276	0.616
Gdp	.1106029	0.001	.1171257	0.001	.1002286	0.017
renew	-.5835462	0.000	-.5542236	0.001	-.5894321	0.000
C	-.3102884	0.432	-.3604896	0.454	-.899405	0.648

The results of AMG estimator are reported in Table 6 for Model 1, Model 2 and Model 3. According to the results of Model 1, the impact of AfT inflows on the energy sector is statistically significant and negative. Thus, AfT has a direct effect on the environmental quality and the increase in AfT inflows to the energy sector leads to increase in environmental quality. This evidence for AfT for the energy sector is supported by the results of Bhattacharyya et al. (2016) indicating that energy aids helped to reduce CO<sub>2</sub> emissions especially in Europe and Central Asian countries, while differed from the findings of Mahalik et al. (2021) demonstrating that energy aids increased CO<sub>2</sub> emissions for India. On the other hand, AfT inflows for other sector have no significant effect on CO<sub>2</sub> emissions. Therefore, it can be concluded that AfT does not have an indirect effect on CO<sub>2</sub> emissions. Among the other explanatory variables, renewable energy consumption reduces CO<sub>2</sub> emissions while economic growth increases them as expected from theory.

According to the results of Model 2, AfT for the renewable energy sector affects CO<sub>2</sub> emissions negatively at a 1% significance level. The results obtained for AfT inflows to the renewable energy sector are consistent with the results of Kablan and Chouard (2022). However, the impact of AfT funds on CO<sub>2</sub> emissions for other sectors is insignificant which is in line with Model 1. Additionally, renewable energy

consumption and economic growth have negative and significantly positive effects on CO<sub>2</sub> emissions, respectively. In summary, the results of Model 1 and Model 2 reveal that AfT has a direct effect but not indirect effect on the environment.

According to Model 3 the effect of total AfT on CO<sub>2</sub> emissions is insignificant. This result differs from the Nguyen et al. (2023) study, which covered 100 countries and found that total AfT had a detrimental effect on environmental quality. The approach that AfT may have an inverted U or U-shaped effect on the environment is not valid for middle income countries. Renewable energy consumption reduces CO<sub>2</sub> emissions while economic growth increases CO<sub>2</sub> emissions in accordance with theory.

Finally, the causal relationships between the variables were also questioned. For this purpose, we employed the causality test developed by Emirmahmutoğlu and Köse (2011). This test can be used for cases where the series are cointegrated as well as not. Another advantage is that it allows series to be integrated of order 0 as well as 1. Moreover, this test considers heterogeneity (Emirmahmutoğlu and Köse, 2011: 872). In this test, the authors adopt the approach of the Toda and Yamamoto (1995) test and produce Fisher statistics for the entire panel, based on the probabilities calculated for each cross-section. The null hypothesis states that there are no causalities for all units, while the alternative hypothesis denotes that there are causalities for some units (Emirmahmutoğlu and Köse, 2011).

According to the causality test results in Table 7, the null hypothesis that AfT for the renewable sector is not a Granger cause of CO<sub>2</sub> emission is rejected at the 5% significance level. Similarly, CO<sub>2</sub> is a Granger cause of AfT funds for the energy sector. However, another finding is that there is no causality running from AfT for other sectors to CO<sub>2</sub> emissions, while it is seen that CO<sub>2</sub> emissions is a Granger cause of AfT funds for all sectors except the energy sectors. There is also a bidirectional causal relationship between AfT funds transferred to the renewable energy sector and CO<sub>2</sub> emissions, while there is a unidirectional causality relationship between AfT funds for other sectors and CO<sub>2</sub> emissions running from AfT funds transferred to other sectors to CO<sub>2</sub> emissions. To summarize, there are causalities from variables reflecting the direct impact of AfT to CO<sub>2</sub> emissions only but not from variables reflecting indirect impact of AfT to CO<sub>2</sub>

emissions. These findings reveal that only AfT funds given to the energy and renewable energy sectors may influence CO<sub>2</sub> emissions and support the evidence obtained for Model 1 and Model 2 in the AMG estimation results.

**Table 7. Results of causality tests**

Null Hypothesis	Fisher stat.	p-value
CO <sub>2</sub> is not Granger cause of AfT <sub>ener</sub>	66.80765	<0.001
AfT <sub>ener</sub> is not Granger cause of CO <sub>2</sub>	37.55394	0.038
CO <sub>2</sub> is not Granger cause of AfT <sub>other1</sub>	69.27143	<0.001
AfT <sub>other1</sub> , is not Granger cause of CO <sub>2</sub>	15.24730	0.913
CO <sub>2</sub> is not Granger cause of AfT <sub>ren</sub>	152.2413	<0.001
AfT <sub>ren</sub> is not Granger cause of CO <sub>2</sub>	120.2964	<0.001
CO <sub>2</sub> is not Granger cause of AfT <sub>other2</sub>	37.30341	0.041
AfT <sub>other2</sub> is not Granger cause of CO <sub>2</sub>	19.81088	0.707
CO <sub>2</sub> is not Granger cause of AfT	41.57285	0.014
AfT is not Granger cause of CO <sub>2</sub>	27.25110	0.293
CO <sub>2</sub> is not Granger cause of AfT <sup>2</sup>	40.37436	0.020
AfT <sup>2</sup> is not Granger cause of CO <sub>2</sub>	26.89424	0.309
CO <sub>2</sub> is not Granger cause of gdp	23.49229	0.490
gdp is not Granger cause of CO <sub>2</sub>	46.21355	0.004
CO <sub>2</sub> is not Granger cause of renew	31.49969	0.140
renew is not Granger cause of CO <sub>2</sub>	38.55326	0.030

The causal relationship between CO<sub>2</sub> emission and total AfT is unidirectional and runs from CO<sub>2</sub> emission to AfT. No causal relationship was found between CO<sub>2</sub> and AfT<sup>2</sup>. These findings are also consistent with the results of the AMG estimator for Model 3. According to Table 7, economic growth and renewable energy consumption have unidirectional relationships with CO<sub>2</sub> emissions. The direction of the causality is from economic growth to CO<sub>2</sub> emissions and from renewable energy consumption to CO<sub>2</sub> emissions, being in line with theory.

## 6. Conclusion

In this study, we investigated the effect of AfT on the environmental quality by focusing on the components of AfT funds in terms of sectors. In this context, we made a distinction between AfT funds for the energy sector and AfT funds for other sectors in

Model 1. By doing so, we aimed to investigate the direct and indirect effect of AfT on CO<sub>2</sub> emissions. Similarly, we also separated the AfT funds into 2 groups as AfT for renewable energy sector and AfT for other sectors to search the direct and indirect impact of AfT in Model 2. Additionally, we analysed the validity of inverted-U-shaped effects between total AfT funds and CO<sub>2</sub> emissions in Model 3.

In our study, a sample consisting of middle-income countries was analysed over the period 2003-2020. We included renewable energy consumption and economic growth variables to all models as control variables. For the coefficient estimates of the regression models, we used the AMG method. According to the AMG results, the direct impacts of AfT on CO<sub>2</sub> emissions are negative and significant for both variables. In contrast, the indirect effects of AfT funds on CO<sub>2</sub> emissions are statistically insignificant revealing that AfT funds for all sectors except energy and renewable energy sectors have no influence on CO<sub>2</sub> emissions.

The AMG results also demonstrate that the impact of total AfT on CO<sub>2</sub> emissions is insignificant (Model 3). Besides, the hypothesis that AfT may have an inverted-U-shaped effect on CO<sub>2</sub> emissions is not confirmed. On the other hand, as expected from theory, renewable energy consumption affects CO<sub>2</sub> emissions negatively, while economic growth affects CO<sub>2</sub> emissions positively in all 3 models.

Finally, we applied the causality test of Emirmahmutoglu and Köse (2011). According to these results, AfT inflows to the renewable energy sector have bidirectional causal relationships with CO<sub>2</sub> emissions. Similarly, AfT funds to the total energy sector have bidirectional causality relationships with CO<sub>2</sub> emissions. No causalities are observed from AfT funds for all other sectors to CO<sub>2</sub> emissions. Moreover, no causality was detected from total AfT to CO<sub>2</sub> emissions. On the other hand, there is a unidirectional causal relationship running from economic growth to CO<sub>2</sub> emissions and from renewable energy consumption to CO<sub>2</sub> emissions. In sum, these findings are in line with the regression model estimates.

To conclude, while AfT has a direct and negative impact on CO<sub>2</sub> emissions, no indirect effect has been observed in middle-income countries. In other words, when we examine the impacts of the components of AfT on CO<sub>2</sub> emissions, only AfT for total energy sector

or renewable energy sector contribute to the reduction of CO<sub>2</sub> emissions. This reveals that AfT funds for the total energy sector are mostly used in renewable energy sector investments. On the other hand, we can conclude that since AfT for other sectors have no effect on CO<sub>2</sub> emissions, AfT has no impact on CO<sub>2</sub> emissions through channels such as economic growth, export diversification, poverty reduction etc. This may result since the countries in the model have realized economic growth and building of infrastructure at a certain level. Besides, it may also take longer for the indirect effect to realize as stated by Kim (2019:2718). On the other hand, the invalidity of the hypothesis that AfT funds may have a U or inverted-U-shaped impact on the environment may be due to the short period analysed.

In summary, the total AfT funds for developing countries to benefit more from foreign trade do not affect environmental quality. However, when we analyse the components of AfT funds on a sectoral basis, AfT funds transferred to renewable energy or to the total energy sector may improve the environmental quality. The nonexistence of the impact of total AfT on CO<sub>2</sub> emissions may occur due to the low share of AfT given to renewable energy or the total energy sector in total AfT. From a climate mitigation policy perspective, AfT donor countries, which care about environmental quality, should increase the proportion of AfT funds for renewable energy and total energy sector. In addition, authorities should encourage renewable energy investments instead of fossil resources to minimize the damage to the environment in middle countries where economic growth is one of the main goals.

## References

- Arvin, M., Dabir-Alai, P. and Lew, B. (2006). Does foreign aid affect the environment in developing economies? *Journal of Economic Development*, 63 (31). 31 (1): 63–87.
- Bhattacharyya, S., M. Intartaglia, and A. McKay 2016. "Does Climate Aid Affect Emissions? Evidence from a Global Dataset." CSAE Working Paper Series 2016-09. Centre for the Study of African Economies, University of Oxford.
- Boly, M. 2018. "CO<sub>2</sub> Mitigation in Developing Countries: The Role of Foreign Aid." Working Papers 201208, CERDI. [http://cerdi.org/production/show/id/1915/type\\_production\\_id/1](http://cerdi.org/production/show/id/1915/type_production_id/1)
- Breusch, T.S & Pagan, A.R. (1980), The Lagrange Multiplier Test and Its Applications to Model Specification Tests in Econometrics. *Review of Economic Studies*, 47, 239-53. <https://doi.org/10.2307/2297111>

- Eberhardt, M. & Bond, S. (2009). Cross-section dependence in nonstationary panel models: A novel estimator. MPRA Paper 17692, University Library of Munich, MPRA Paper, No. 17870. [http://mpra.ub.uni-muenchen.de/17692/1/MPRA\\_paper\\_17692.pdf](http://mpra.ub.uni-muenchen.de/17692/1/MPRA_paper_17692.pdf).
- Eberhardt, M. (2012) Estimating panel time-series models with heterogeneous slopes, *The Stata Journal* (2012), 12, Number 1, pp. 61–71. <https://doi.org/10.1177/1536867X1201200105>
- Eberhardt, M. & Teal, F. (2010) Productivity analysis in global manufacturing production. Discussion Paper 515, Department of Economics, University of Oxford. <http://www.economics.ox.ac.uk/research/WP/pdf/paper515.pdf>.
- Emirmahmutoglu, F., & Köse, N. (2011). Testing for granger causality in heterogeneous mixed panels. *Economic Modelling*, 28, 870–876. <https://doi.org/10.1016/j.econmod.2010.10.018>
- Kablan, S., & Chouard, V. (2022) Does climate aid matter for reducing CO2 emissions? The case of foreign aid for renewable energy, *Applied Economics*, 54:46,5357-5372, <https://doi.org/10.1080/00036846.2022.2044995>.
- Kim, Y. R.,(2019). Does aid for trade diversify the export structure of recipient countries? *The World Economy*, 42(9), 2684–2722. <https://doi.org/10.1111/twec.12845>.
- Mahalik MK, Villanthenkodath MA, Mallick H, Gupta M (2021) Assessing the effectiveness of total foreign aid and foreign energy aid inflows on environmental quality in India. *Energy Policy*, 149(112015), 1-19. <https://doi.org/10.1016/j.enpol.2020.112015>.
- Nguyen, T. P. T., Nguyen, V. A., & Ly-My, D. (2023). Does Aid for Trade affect the quality of the environment? Evidence from Aid for Trade recipient countries. *Journal of International Development*, 1–20. <https://doi.org/10.1002/jid.3780>.
- OECD/WTO (2022), Aid for Trade at a Glance 2022: Empowering Connected, Sustainable Trade, OECD Publishing, Paris, <https://doi.org/10.1787/9ce2b7ba-en>.
- OECD (2024), OECD.Stat, Creditor Reporting System, <https://stats.oecd.org/>.
- Pesaran, M. H. (2007), A Simple Panel Unit Root Test in the Presence of Cross Section Dependence, *Journal of Applied Econometrics*, 22(2), pp. 265-312. <https://doi.org/10.1002/jae.951>
- Pesaran, M. H., Yamagata, T. (2008) Testing slope homogeneity in large panels, *Journal of Econometrics*, 142(1), pp. 50–93. <https://doi.org/10.1016/j.jeconom.2007.05.010>
- Pesaran, M.H., Ullah, A. and Yamagata, T. (2008), “A Bias-Adjusted LM Test of Error Cross-Section Independence”, *Econometrics Journal*, 11(1), 105-127. <https://doi.org/10.1111/j.1368-423X.2007.00227.x>
- Toda, H.Y., Yamamoto, T., 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* 66, 225–250. [https://doi.org/10.1016/0304-4076\(94\)01616-8](https://doi.org/10.1016/0304-4076(94)01616-8)
- Westerlund, J. (2008). Panel Cointegration Tests of the Fisher Effect. *Journal of Applied Econometrics* 23(2), 193-223. <https://doi.org/10.1002/jae.967>

# Fra prompt til praksis

## Undervisning i anvendt statistik med generativ AI

Sara Armandi, Hans Bay, Anders Milhøj, Markus Roed Schøler og Nina Johanna Åberg-Jensen  
Økonomisk Institut, Københavns Universitet

***Disclaimer:** Artikel er skrevet i en tid med hastig udvikling inden for generativ AI, og flere oplysninger kan allerede ved udgivelsen være forældede. Dette understreger vigtigheden af kontinuerlig opmærksomhed og forskning inden for området.*

### **Abstract**

Denne artikel undersøger integrationen af generativ AI, med særligt fokus på store sprogmodeller som ChatGPT og Copilot, i undervisningen i anvendt statistik. Fokus er rettet mod sprogmodellernes potentiale til at understøtte læring i kodning og dataanalyse, særligt med henblik på programmet SAS. Gennem praktiske eksempler fremhæves både styrker og begrænsninger ved brugen af modellerne, herunder vigtigheden af velstruktureret promptning for at opnå de bedste resultater. Artiklen adresserer også didaktiske overvejelser, der kan hjælpe undervisere og studerende med at udnytte modellerne effektivt. Afslutningsvis præsenteres anbefalinger til, hvordan undervisnings- og eksamensformer kan tilpasses for at fremme lige læringsvilkår og sikre en balanceret anvendelse af AI i den akademiske kontekst.

### **1 Introduktion**

Generativ AI (GenAI) er blevet en integreret del af vores hverdag, især gennem store sprogmodeller (LLM'er, "Large Language Models") som OpenAI's ChatGPT og Microsofts



Copilot. Disse modeller kan både generere tekst og assistere med opgaver inden for programmering og dataanalyse.

Pr. 2024 er der empirisk belæg for, at ChatGPT er den mest udbredte sprogmodel blandt studerende på Økonomistudiet ved Københavns Universitet, der tager kurser i anvendt statistik. Universitetet har desuden anskaffet en Enterprise-licens til Copilot, hvilket gør dette værktøj frit tilgængeligt for både ansatte og studerende. Mange studerende er dog usikre på, hvordan de korrekt og hensigtsmæssigt kan anvende disse værktøjer, hvilket bekræftes af en større spørgeskemaundersøgelse fra de humanistiske og teologiske fakulteter. I flere kurser frarådes studerende decideret at bruge GenAI-værktøjer, muligvis fordi underviserne ikke har overblik over konsekvenserne ved anvendelsen.

Der er et presserende behov for undervisning i brugen af GenAI-værktøjer, da mange undervisere oplever at modtage opgaver, der tydeligt viser sig at være delvist genereret af LLM'er. Samtidig skaber forskelle i studerendes evner til at anvende modellerne en uligevægt i forudsætningerne for læring og resultater, hvilket gør det afgørende at integrere undervisning i LLM'er for at fremme lige vilkår for alle studerende.

Det er vigtigt at bemærke, at der ikke kun er behov for undervisning af studerende; mange undervisere har også brug for en dybere forståelse af GenAI. Behovet for at dykke ned i anvendelsen af GenAI understreges af følgende officielle udmelding fra Københavns Universitet:

*"Fra og med efteråret 2025 vil anvendelsen af generativ AI (GAI) ved alle eksamensformer (inkl. hjemmeopgaver, BA-projekter og specialer) med "alle hjælpemidler tilladt" derfor som standard være tilladt på KU."*

Udfordringen ligger i, at de nuværende undervisnings-, opgave- og eksamensformer ikke nødvendigvis er designet til at håndtere GenAI, da det, jf. Figur 1, bl.a. er muligt for studerende at opnå tilfredsstillende resultater ved blot at benytte generiske svar fra LLM'er uden at engagere sig i den nødvendige læringsproces.

Figur 1: Eksempel på korrespondance med ChatGPT



Kilde: Eksempel fra ChatGPT.

Hvis der tages højde for de nævnte udfordringer, har GenAI-teknologier potentiale til at revolutionere undervisningen, herunder inden for områder som programmering og anvendt statistik. LLM'erne kan understøtte udviklingen af kode samt forbedre forståelsen og fortolkningen af dataanalyse. Denne artikel undersøger, hvordan LLM'er kan integreres i universitetsundervisningen med fokus på anvendt statistik, især i relation til programmerings- og kodeopgaver til analyser. Der tages udgangspunkt i to specifikke generative AI-værktøjer: ChatGPT og Copilot, som begge er blevet afprøvet i forbindelse med opgaveløsning. Gennem praktiske erfaringer belyses, hvordan disse værktøjer kan implementeres effektivt i undervisningen.

## 2 Hvordan virker LLM'erne?

GenAI er en gren af AI, der fokuserer på at skabe nyt indhold eller data ved at lære fra eksisterende data, såsom tekst, kode, video og billeder. LLM'er er avancerede GenAI-

modeller, der er designet til at forstå og generere menneskelignende tekst baseret på enorme mængder tekstdata, hvilket inkluderer information fra internettet, bøger, artikler og andre kilder. En LLM er en kraftfuld machine learning model, som kan bearbejde og identificere komplekse sammenhænge i tekster. Disse modeller bygger på teknikker som dybe neutrale netværk, som er designet til at lære komplekse mønstre og relationer i data gennem hierarkisk repræsentation. Alt dette resulterer i modeller med op til milliarder af parametre.

Modellerne fungerer ved at bearbejde tekst i segmenter snarere end ord for ord. Disse segmenter behandles parallelt, hvilket gør det muligt for modellen at forstå komplekse sammenhænge og mønstre i teksten. Ved at analysere flere segmenter samtidigt kan modellen forudsige det næste ord eller sætning baseret på den givne kontekst.

Helt konkret indebærer det, at modellen, baseret på et bagvedliggende system af unikke ordrepræsentationer, kan knytte ord til vektorer i et højdimensionalt rum. Dette muliggør beregningen af sandsynligheder for disse repræsentationer og dermed sandsynligheden for et givent ord. Selvom sandsynligheden for et bestemt ord kan være høj, er der altså stadig en sandsynlighed for, at et andet ord vælges, jf. Figur 2. Denne egenskab ved modellerne gør det muligt for dem at generere nyt og unikt indhold, herunder sammenhængende sætninger og tekst.

Figur 2: Ordforudsigelser og sandsynligheder fra en LLM



Kilde: Eksempel fra ChatGPT Playground

Det er vigtigt at bemærke, at LLM'er som udgangspunkt ikke er matematiske modeller designet til at udføre beregninger eller anvende formler, hvilket kan gøre dem upålidelige til præcise numeriske opgaver og avanceret logik. Derudover kan de have svært ved at forstå komplekse kontekster eller subtiliteter, hvilket kan føre til misforståelser eller upræcise svar. Disse begrænsninger understreger, at LLM'er, på trods af deres evne til at generere tekst og analysere sprog, bør bruges med omtanke – især i fag som anvendt statistik, hvor nøjagtighed og præcision er afgørende for resultaterne.

## **2.1 Ikke længere kun en sprogmodel**

Selvom LLM'er primært er baseret på sprog, har den hurtige udvikling ført til, at flere modeller nu også kan håndtere matematiske og kodningsrelaterede opgaver. Dette opnås gennem en kombination af avanceret mønstergenkendelse og evnen til at generere samt eksekvere kode, hvilket gør dem særligt nyttige i undervisningssammenhænge inden for anvendt statistik.

Når en LLM modtager en matematisk forespørgsel, analyserer den først strukturen og kravene for at identificere relevante regler eller mønstre. Enkle problemer løses typisk ved hjælp af mønstre fra LLM'ens træningsdata, mens mere komplekse opgaver kræver, at modellen genererer kode, som derefter eksekveres for at sikre korrektheden. Denne proces muliggøres gennem en kodefortolker, der omsætter tekstbaserede beskrivelser til faktiske beregninger og kodning.

En kodefortolker, der ofte er integreret med Python, giver modellen mulighed for at oversætte forespørgsler til funktionelle, kodebaserede løsninger, der kan udføres direkte. Dette muliggør, at modellen kan håndtere alt fra præcise beregninger og avancerede dataanalyser til generering af grafer og visualiseringer. Resultaterne præsenteres på en overskuelig måde og suppleres ofte med forklaringer af fremgangsmåden. Eksempler på denne funktionalitet gennemgås kort i Bilag 1.

### **3 Teknisk funktionalitet og begrænsninger af LLM'er**

Moderne LLM'er som ChatGPT og Copilot tilbyder en bred vifte af anvendelser inden for uddannelse, men de har også tekniske begrænsninger og særlige funktioner, der bør tages i betragtning. En af de praktiske funktioner er muligheden for at arbejde med hukommelse på tværs af samtaler. Dette kan være nyttigt ved større projekter, hvor modellen lærer brugerens præferencer og arbejdsmetoder. Det betyder, at man kan angive detaljer om data eller opgavekrav én gang og bruge denne kontekst i senere prompts. Ulempen er dog, at denne hukommelse kræver vedligeholdelse; hvis man ikke opdaterer modellen om ændringer i projektet, kan dens svar blive mindre relevante – det er som at arbejde med en kollega, man glemmer at informere. Ved kortere opgaver eller isolerede projekter kan det være en fordel at arbejde uden hukommelse på tværs af samtaler. Ved at starte en ny samtale for hvert projekt sikrer man, at tidligere viden ikke påvirker modellens output.

#### **3.1 Træningsdataenes betydning**

LLM'er præsterer ofte bedst på engelsk, da de engelske modeller er trænet på større datamængder og derfor har mere viden til rådighed. Dog kan der være situationer, hvor modellen yder bedre på dansk, især når spørgsmålet drejer sig om lokale emner, der primært er beskrevet i danske sammenhænge.

Inden for SAS-kodning er det tydeligt, at LLM'er er bedre trænet i PROC SQL end i SAS' data steps. Dette skyldes, at SQL er et mere udbredt programmeringssprog, der anvendes på tværs af forskellige databaser. Derfor genererer LLM'er oftere kode i PROC SQL fremfor kode i SAS data steps, når en opgave kan løses ved begge metoder. Dette er en konsekvens af begrænset træningsdata, hvilket også ses for specifikke SAS-procedurer.

#### **3.2 Forskelle mellem gratis- og betalingsversioner af LLM'er**

En væsentlig forskel mellem gratis- og betalingsversioner af LLM'er er deres evne til at håndtere større datamængder. Generelt er både antallet og længden af prompts begrænset i gratisversionerne. Tabel 1 viser forskellene i pris samt begrænsningerne for filupload for

Copilot Enterprise og de nuværende tre versioner af ChatGPT, som private brugere kan anvende. Københavns Universitet har betalt for en Enterprise-licens til Copilot, så selvom studerende og undervisere ikke selv betaler, er licensen ikke gratis. Tabellen giver derfor et overblik over de udgifter, som studerende og undervisere ved Københavns Universitet vil have i forbindelse med brugen af disse værktøjer.

Tabel 1: KU-brugerpris og filupload for hhv. Copilot Enterprise og ChatGPT

	<b>Copilot Enterprise</b>	<b>ChatGPT Gratis</b>	<b>ChatGPT Plus</b>	<b>ChatGPT Pro</b>
<b>Pris pr. måned</b>	\$0 USD/måned	\$0 USD/måned	\$20 USD/måned	\$200 USD/måned
	<b>Muligheder ved upload af filer</b>			
<b>Filstørrelse</b>				
- CSV	10 MB pr. fil	50 MB	50 MB	50 MB
- Billeder	Håndterer ikke billeder	20 MB	20 MB	20 MB
- Tekstfiler	10 MB pr. fil	20 MB	20 MB	20 MB
<b>Tilladt antal uploadede filer</b>	Ubegrænset antal, maks. 2 GB	3 filer dagligt, maks. 10 GB	80 filer hver tredje time, maks. 10 GB	80 filer hver tredje time, maks. 10 GB

Kilde: OpenAI Help Center, <https://help.openai.com/en/articles/8555545-file-uploads-faq> og <https://copilot.cloud.microsoft/>, december 2024

Hvis en fil uploades i en samtale, og denne når sin længdegrænse, kan det blive vanskeligt at fortsætte interaktionen. Dette kan resultere i tab af uploadet information og en følelse af at skulle starte forfra.

Som det fremgår af Tabel 1 understøtter betalingsversioner mere data, og dermed mere komplekse opgaver. For eksempel har ChatGPT's Plus-version vist sig at kunne løse avancerede statistiske opgaver, som besvarelsen af en SAS-opgave i kurset *Videregående Statistik*. Her genererede modellen komplette svar, inklusive kode og tabeller, uden at skulle åbne SAS-editoren. En begrænsning opstod dog, idet datasættet i SAS-format skulle konverteres til et format, som modellen kunne genkende, hvilket LLM'en dog kunne løse ved at generere noget kode.

Når data og opgavebeskrivelser uploades, kan LLM'er generere komplette løsninger, herunder figurer, statistiske modeller og fortolkninger af resultater. Det er dog vigtigt at bemærke, at da modellerne anvender Python, kan der opstå små diskrepanser mellem udregninger udført direkte i modellerne og dem udført i SAS. Desuden vil output fra modellerne være formateret anderledes end for eksempel SAS-output.

Udviklingen af LLM'er går hurtigt, og hvad der i dag kræver en betalingsversion, kan snart blive en del af den gratis udgave.

## **4 Prompt Engineering for studerende**

En prompt er en tekstuel instruktion, der guider en LLM i, hvilken type output eller indhold den skal generere. Promptudvikling er kunsten at formulere klare og præcise instruktioner, der optimerer modellens ydeevne og hjælper brugerne med at maksimere deres interaktion med LLM'er. Denne disciplin er i konstant udvikling, da metoderne til at skabe effektive instruktioner hurtigt tilpasser sig nye behov og teknologiske fremskridt.

### **4.1 Prompt-opskrift**

Der er mange forskellige måder at prompte på, og mange forskellige tilgange afhængig af opgavetyper. Tabel 2 beskriver kort de vigtigste elementer, der bør overvejes i opbygningen af en prompt, når der arbejdes med anvendt statistik – især i relation til kode og fortolkninger af analytisk output.



Tabel 2: Elementer i prompt til effektiv anvendelse af LLM'er

	Element	Beskrivelse	Eksempel
————— Vigtigt ————— ————— Valgfrit —————	Opgave	Giv LLM'en en konkret opgave med tydeligt mål. Start med handlingsverber som: "Generér", "Skriv", "Analysér"	"Generér en SAS-kode, der [beskriv opgaven]"
	Kontekst	Inkluder relevante detaljer for at optimere output. Tænk på hv-ord: "Hvem", "Hvad", "Hvornår", "Hvor", "Hvorfor", "Hvordan". Tidligere kørt kode kan være essentielt.	"Koden skal anvendes til at analysere [indsæt datasæt]"
	Eksempel	Brug eksempler for at forbedre kvalitet og præcision af output. Det er ofte lettere at vise, hvad man ønsker, end at beskrive det.	"En tidligere SAS-kode ser således ud: [indsæt kode]"
	Persona	Beskriv specifik ekspertise eller baggrund for at tilpasse output-formatet. Brug "Ager som [persona]".	"Du er dataanalytiker med erfaring i SAS..."
	Format	Angiv bestemt format for at gøre output struktureret og forståeligt. Formatér inkluderer punktopstillinger, kodeblokke, paragraffer eller tabeller.	"Formater koden så den let kopieres ind i en SAS-editor"

Kilde: Mange, men især <https://www.syv.ai/prompting-guide>.

Selvom en omfattende kontekst ofte forbedrer kvaliteten af output fra LLM'er, kan det i enklere opgaver være en fordel at begrænse den givne kontekst. Stramme rammer kan nemlig hæmme LLM'ernes evne til at generere det mest fordelagtige output, især når en opgave kan løses på flere måder, og metoden er mindre vigtig. I tilfælde af simple opgaver kan det derfor være mere effektivt at undlade kodeeksempler og i stedet formulere en kort og præcis prompt, der klart angiver det ønskede output. Dette gælder også, når man ønsker at forstå specifik kode eller output. Overdreven forklaring af egne tanker kan begrænse LLM'ernes muligheder og føre til suboptimale resultater.

Der findes flere interessante promptelementer og -teknikker, der kan være nyttige inden for anvendt statistik. For eksempel kan man anmode om flere forslag til en opgave for at vælge det mest optimale resultat. Desuden kan det være gavnligt at bede LLM'erne om at stille spørgsmål til sin prompt, hvilket kan resultere i en mere præcis prompt og et mere nøjagtigt output.

## 4.2 Simpel vs. avanceret promptning

Med udgangspunkt i eksamensopgaven i faget *Programmering og Statistik med SAS* fra juni 2021 undersøges hvordan forskellige måder at prompte på kan have afgørende indflydelse på det output, der opnås. Eksamensopgaven var en 3½-timers hjemmearbejdsopgave udarbejdet af Anders Milhøj. Idet opgaven blev stillet før udbredelsen af GenAI, tog den ikke højde for muligheden for at anvende GenAI til løsning af opgaverne.

Gennem flere itereringer er det blevet undersøgt, hvordan forskellige prompts kan hjælpe med besvarelsen af eksamensopgaven. I denne artikel præsenteres to yderpunkter, der betegnes som henholdsvis simpel promptning og avanceret promptning. Studerende, der anvender simpel promptning, er ofte dem, der ikke har fået uddannelse i brugen af LLM'er eller dem, der er fagligt udfordret. På den anden side kendetegnes avanceret promptning ofte ved, at den anvendes af fagligt dygtige og grundige studerende, der har en god forståelse af pensum.

Tabel 3 viser fremgangsmåden samt de umiddelbare resultater for de to forskellige tilgange til at besvare eksamensopgaven. Det er tydeligt, at mere avancerede promptteknikker resulterer i langt mere brugbare og fornuftige resultater. Ved at anvende avancerede prompts får LLM'en mulighed for at producere mere korrekt og relevant output, da flere kontekstuelle elementer er inkluderet. Dette understreger behovet for didaktisk støtte i promptkonstruktion, som kan guide mindre erfarne studerende.

Ud fra Tabel 3 ses også hvordan forskelle i promptstrategi påvirker kvaliteten af output og den læringsmæssige værdi. Det fremgår, at manglende uddannelse i brugen af LLM'er i undervisningen kan skabe ulighed blandt studerende. Derfor er det vigtigt at tilbyde uddannelse i brugen af LLM'er for at hjælpe studerende, der ikke har kendskab til effektive promptteknikker.

Tabel 3: Forskel på hhv. simpel- og avanceret promptning

	<i>Promptning</i>	<i>Simpel promptning</i>	<i>Avanceret promptning</i>
Input til prompt	Tilgang	En simpel prompt uden detaljeret kontekst	En omhyggeligt opbygget prompt med detaljeret kontekst og relevante prompt-elementer
	Struktur	Indsætter opgaveteksten som direkte prompt uden yderligere detaljer	Starter med en generel prompt og tilføjer (evt. trinvis) mere kontekst og specifik viden
	Eksempel	"Skriv kode til denne opgave i SAS"	"Skriv SAS-kode til dataanalyse med korrekt brug af libname og stier, undgå dataoverskrivning"
Output fra prompt	Kvalitet	Ofte mangelfuldt output, der kræver SAS-viden til fejlsøgning	Ofte præcist og korrekt output i første forsøg
	Fejl og fejlsøgning	Overskriver ofte data uden advarsel; kræver manuel tilpasning	Færre fejl i output; undgår typiske fejl ved datahåndtering og SAS-opsætning
	Anvendelse af output	Afhænger af erfaring med SAS og manuel justering af fejl	Kan direkte anvende output uden omfattende redigering
	Læring og refleksion	Begrænset læring; primært afhængig af, at AI "løser" opgaven	Øget læring; promptdesignet kræver forståelse af opgavens krav og SAS-komponenter

Kilde: Egne observationer

For studerende med begrænset kendskab til pensum kan en praktisk tilgang være at indsætte selve opgaveteksten i LLM'en uden yderligere prompt engineering for at se, hvilket output der genereres. Dette kan føre til korrekte løsninger på mere simple opgaver, såsom visse opgaver i *Programmering og Statistik med SAS*, som hidtil har været formuleret. Dog kan studerende opnå markant bedre resultater, hvis de har en grundlæggende forståelse af,

hvordan LLM'erne fungerer, samt hvilke elementer der gør en prompt effektiv. Selv en grundlæggende viden om disse aspekter kan hjælpe dem med at formulere bedre prompts, hvilket i sidste ende kan føre til mere præcise og relevante output samt forbedrede opgavebesvarelser.

### 4.3 Den gode prompt til generering af SAS-kode

Med udgangspunkt i en konkret opgave fra eksamenssættet i kurset *Programmering og Statistik med SAS* fra juni 2021, undersøges, hvordan man kan opnå det bedst mulige output ved at anvende forskellige prompts til at generere den nødvendige SAS-kode for at besvare følgende eksamensopgave:

*2a) Træk sognenummeret ud af variabelen sogn i datasættet sogne\_udd\_kirke.sas7bdat i en numerisk variabel kaldet sognenummer. Brug både datasteps og SQL.*

Når opgaverne indeholder mere tekst, er det ofte nødvendigt at forenkle dem væsentligt for at hjælpe LLM'erne med at løse dem. Den konkrete opgave er kortfattet formuleret, hvilket gør det klart, hvad der skal gøres, og dermed giver LLM'erne gode betingelser for at løse opgaven.

Ved brug af gratisversionen af ChatGPT, Model 3.5/4 den 21. oktober 2024, startes med en simpel prompt, hvor der udover opgaveteksten tilføjes:

*"Jeg arbejder i dataprogrammet SAS. Generér en kode til følgende opgave:"*

Den genererede kode, som ses i Tabel 4, formår at identificere sognenummeret korrekt. Der opstår dog et problem idet det oprindelige datasæt overskrives, hvilket kan føre til problemer for de studerende i efterfølgende opgaver. Derudover bliver der ikke dannet noget udskrift, hvilket kan gøre det vanskeligt for de studerende at vurdere, om opgaven er løst korrekt.

For at sikre, at svarene fra LLM'en ikke påvirkes af tidligere stillede spørgsmål og prompts, skrives de forskellige prompts i separate samtaler uden hukommelse mellem dem. Dette betyder, at der ikke er nogen historik i samtalerne. Ved at anvende elementer beskrevet i Tabel 2 fra Afsnit 4.1, udvides prompten til at inkludere flere relevante detaljer. Forståelsen af eksamensopgaven og de mulige løsningsmetoder i SAS er en vigtig del af promptdesignet, da de studerende selv skal kunne ræsonnere sig frem til, at der er behov for at oprette et nyt datasæt.

Tabel 4: Kode genereret af ChatGPT ved hhv. simpel og endelig prompt

<i>Simpel prompt</i>	<i>Endelig prompt</i>
<pre>data sogne_udd_kirke;   set sogne_udd_kirke;   sognenummer = input(substr(sogn, 1, 4), 8.); /* Ekstraher de første 4 tegn fra 'sogn' og konverter til numerisk */ run;</pre>	<pre>libname sasprog 'path'; /* Ensure the library is defined */ data sognenummer_data; /* New dataset name */   set sasprog.sogne_udd_kirke; /* Use the original dataset */   /* Extract the first 4 characters as a numerical variable */   sognenummer = input(substr(sogn, 1, 4), 8.); /* Convert the extracted substring to a number */ run; /* Print the data to verify */ proc print data=sognenummer_data (obs=10); /* Print the first 10 observations */   title "First 10 Observations of the New Dataset sognenummer_data"; run;</pre>

Kilde: Output fra ChatGPT, Model 3.5/4, oktober 2024

Den endelige prompt formuleres på engelsk, da ChatGPT har vist sig at være mere effektiv til at generere SAS-kode, når der prompts på engelsk. Den endelige prompt lyder således:

*“I am an economics student working on an assignment in the computer program SAS. [Kontekst] The data set I'm working with is called sogne\_udd\_kirke and contains information regarding the population's educational level in different counties in Denmark. [Opgave] Please generate a SAS code to solve the following: Create a new data set where the county number is extracted from the variable sogn and made into a numerical variable called sognenummer. [Eksempel] The county*

*number is the first 4 characters in the variable sogn, and the number and the name of the county are separated by a blank. [Format] I would like the code to include both data steps and SQL. Additionally, please include a print statement to check that the code is correct.”*

I ovenstående eksempel bidrager den første linje til formateringen ved at fremhæve et vigtigt aspekt: at være opmærksom på modtageren af en prompt.

Samme forsøg udføres i Copilot Enterprise, som genererer tilsvarende resultater. Den største forskel ligger i, at Copilot ikke kommer med alle de pædagogiske forklaringer, som ChatGPT gør, uanset om man siger, at man er studerende eller ej. Desuden printer ChatGPT ofte kodeforklaringer for hver linje, som kan stå i SAS. Svaret på hhv. den simple og den endelige prompt ses i Tabel 4. I tabellen ses kun den del af svaret, der inkluderer koden til datasteps eftersom SQL-koden er løst på tilsvarende vis.

#### **4.4 Systemprompt**

Der er forskel på output fra ChatGPT og Copilot, da der er forskel på systemprompten. En systemprompt er en tekst eller instruktion, der gives til en LLM for helt overordnet, og på tværs af samtaler, at guide dens svar eller adfærd. Den hjælper med at definere konteksten og formålet med interaktionen, hvilket sikrer, at modellen leverer relevante og præcise oplysninger.

Ved sammenligning af gratisversionen af ChatGPT og Copilot Enterprise, kan der ved samme prompt, og uden ændring i lokale brugerindstillinger ses tydelig forskel i output, idet ChatGPT for SAS-kodeeksempler generelt udviser bedre pædagogiske evner i form af forklaringer og formatering end Copilot.

I ChatGPT er det muligt at tilpasse instruktionerne i sin lokale ChatGPT-version. Således kan man påvirke systemprompten ved at tilføje svar til følgende to spørgsmål; ”What would you like ChatGPT to know about you to provide better responses?” og ” How would you like ChatGPT to respond?”. Tabel 5 viser et eksempel på tilpasning af systemprompten som gemmes i de generelle indstillinger.

Tabel 5: Customisering af systemprompt i ChatGPT

<p><b>Custom Instructions</b></p> <p>What would you like ChatGPT to know about you to provide better responses?</p> <p><i>I am an economics student training my skills in the computer program SAS. I have created a library in SAS called SASPROG, meaning that I have already run the code “libname SASPROG ‘path’;” with success.</i></p> <p>How would you like ChatGPT to respond?</p> <p><i>When explaining something, the answer should be short and in bullets. When I ask about a new SAS procedure, I would like a brief explanation of the procedure, what it is used for and with a basic code example. When I ask about code, I would like a short explanation for all the code lines in the code, ready to paste into SAS. All the original datasets I am working with is saved in the library SASPROG, meaning that when pulling data from an original dataset, you should always write SASPROG. Before the name of the dataset. When you create code, I would NOT like you to save the new dataset in the SASPROG library.</i></p>
---

Kilde: System-prompt i ChatGPT

Med den ovenstående systemprompt kan ChatGPT nu løse opgaver, som den tidligere havde problemer med, fejlfrit. Selvom prompten er på engelsk, kan opgaverne løses både på dansk og engelsk. De tidligere udfordringer med at trække data og overskrive datasæt er blevet løst, da prompten klart angiver, hvilket bibliotek data skal hentes fra, og præciserer, at nye datasæt ikke skal gemmes i dette bibliotek.

Selvom det ikke er muligt at oprette en systemprompt direkte i Copilot, kan den indsættes manuelt i starten af en samtale. På den måde genereres kodeoutput af samme kvalitet som hos ChatGPT. Ved længerevarende opgaveløsning af samme type skal man dog huske at indsætte prompten hver gang og i alle samtaler for at opretholde outputkvaliteten.

## 5 Anvendelse af GenAI i undervisningen

I foråret 2025 introduceres det nye kursus *SAS og AI i Praksis*, som erstatter det tidligere *Programmering og Statistik med SAS*. Dette kursus har til formål at integrere GenAI i undervisningen, hvilket åbner op for en mere effektiv læringsmetode. Ved at anvende LLM'er kan undervisningen accelereres, idet grundlæggende programmeringsfærdigheder kan tilegnes på halvdelen af den tidligere tid. Fokus vil være på essentielle koncepter som syntaks, libname, SAS-biblioteker og dataoverskrivning, samt de områder, hvor LLM'erne kan have begrænsninger.

### 5.1 Styrker og Svagheder ved LLM'er i undervisningen

Styrker:

- Effektivt læringsværktøj
- Forklaring af kode og grundlæggende procedurer
- Hjælp til at lære nye muligheder i kald, f.eks. optioner i Proc Sgplot
- Generering af eksempeldata til mere kedelige opgaver
- Automatisering af manuelle processer

Svagheder:

- Vanskeligheder med komplekse matematiske og avancerede logiske opgaver
- Udfordringer med makroer og nye procedurer
- Begrænset viden af meget specifik og konkret kode
- Sammenblanding med andre programmeringssprog
- Fejl i optioner/statements, der lyder plausible, men ikke eksisterer, som f.eks. PROC PYTHON vs. PROC R.



## 5.2 Opgaver før og nu

LLM'er har vist sig at være særdeles effektive til at løse eksamensopgaver, forudsat at der er forståelse for effektiv udarbejdelse af prompter. Dette rejser spørgsmålet om nødvendigheden af at revurdere opgaveformuleringerne. Studerende opfordres til at angive, hvilken LLM de har anvendt til at besvare opgaverne.

Traditionelle opgaver, som har bestået af at finde fejl i koder eller forklare samtlige kodelinjer i et program, har mistet deres relevans, da LLM'erne kan give præcise svar. I stedet bør der fokuseres på projektbaserede opgaver, der fremmer dybere forståelse. Et eksempel kunne være at give studerende en kode, som de skal optimere eller anvende til at løse et konkret problem, som vist i Tabel 6.

Tabel 6: Eksempel på traditionel vs. ny opgave indenfor programmering

<i>Traditionel opgave</i>	<i>Ny opgave</i>
<p>Forklar følgende kode:</p> <pre>data d2; retain studium; set d1; if anydigit(tekst)=0 and index(tekst, 'Adgangskvotient')=0 and   index(tekst, 'Standby- kvotient')=0 and   tekst ne 'Ansoegninger' then studium=tekst; run;</pre>	<ol style="list-style-type: none"><li>1. Hvad gør følgende kode? proc sgplot data=sashelp.class; vbox weight; run;</li><li>2. Rotér boxplottet, så det er horisontalt fremfor vertikalt.</li><li>3. Lav et nyt boxplot ud fra koden i opg. 1, hvor du opdeler populationen efter køn, dvs. at du har to boxplots i samme graf.</li><li>4. Sæt danske navne på akserne til boxplottet genereret i opgave 3.</li><li>5. Find et andet kald til at generere boxplot end proc sgplot og find ud af forskellen mellem procedureerne.</li></ol>

Kilde: Egne observationer

Den traditionelle opgave i Tabel 6 kan uden videre løses af GenAI og sættes ind i en opgavebesvarelse uden at studerende selv forstår hvad koden gør. I den nye opgavetype lærer studerende forskellige muligheder når man genererer boxplot. Derudover bliver der stillet krav til at studerende tester koderne, der eventuelt genereres sammen med GenAI.

Eftersom GenAI endnu ikke kan producere output, der visuelt ligner SAS-output, bør der stilles krav om at vise specifikt output fra SAS-kørsel. LLM'er kan generere visuelle repræsentationer af tabeller og grafer, men ikke endnu genskabe SAS-output i præcist samme format. Studerende kan derfor anmodes om at indsende screenshots fra SAS for at demonstrere deres evne til at producere det ønskede output.

Ved at integrere GenAI i undervisningen ønskes ikke blot at effektivisere læringsprocessen, men også at opgaver og eksamener tilpasses til den moderne teknologiske virkelighed.

## **6 Diskussion og konklusion**

Integrationen af GenAI, som ChatGPT og Copilot, revolutionerer måden, vi underviser. Disse værktøjer tilbyder betydelige fordele ved at frigive tid fra rutineprægede opgaver, men de medfører også risici for de studerendes grundlæggende forståelse af emnet. Det er essentielt, at undervisere fokuserer på at styrke de studerendes kritiske tænkning ved at fremhæve både fordele og begrænsninger ved brugen af LLM'erne.

Selvom AI-modeller kan være nyttige til at guide studerende og løse simple opgaver, er det op til den enkelte studerende at forstå, hvordan man analyserer et problem og formulerer relevante spørgsmål. Uden en solid forståelse vil der være en risiko for, at outputtet fra modellerne bliver utilfredsstillende, et fænomen ofte beskrevet som "garbage in, garbage out." Derfor er det afgørende, at studerende lærer at formulere velovervejede prompts, snarere end at prompte hovedløst.

Desuden er det vigtigt at overveje forskellene mellem versioner og LLM'er, samt omkostningerne ved betalings- versus gratisversioner. Studerende, der investerer i en avanceret LLM, opnår kun en fordel, hvis de kan udnytte den korrekt. En grundig og veluddannet studerende vil ofte opnå bedre resultater med en gratis model end en mindre engageret studerende, der betaler for en bedre version. Med de i skrivende stund fordele ved betalingsversionerne, har studerende dog givet udtryk for, at de i eksamensperioder vil

sikre sig adgang til en betalingsudgave af LLM'erne. Derfor er det vigtigt at sikre, at LLM'erne ikke får for stor indflydelse på (eksamens-)opgaver.

En yderligere udfordring er den hastige udvikling inden for generativ AI, hvor nye funktioner og opdateringer ofte sker uden tydelig kommunikation. Det kan være svært at kvantificere forskellene mellem gamle og nye versioner af samme produkt, hvilket stiller krav til både undervisere og studerende om at holde sig ajour med de nyeste ændringer.

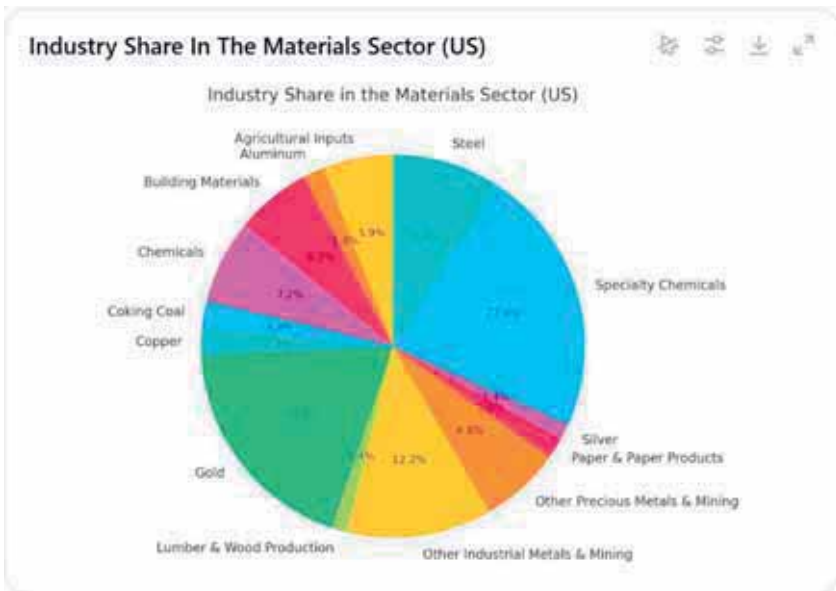
Dette studie fremhæver således både mulighederne og udfordringerne ved at integrere generativ AI i undervisningen i anvendt statistik. Selvom AI-værktøjerne åbner op for nye metoder til at arbejde med kodning og dataanalyse, kræver det en tilpasning af undervisningens opgave- og læringsmål for at sikre optimal læring. Fremtidig forskning vil fokusere på at undersøge effekterne af integrationen af generativ AI i undervisningen i anvendt statistik, hvilket kan bidrage til en dybere forståelse af, hvordan generativ AI kan forbedre læringsprocesser.

## Bilag

### Bilag 1 Output fra ChatGPT's kodeförtolker

Et eksempel på ChatGPT's kodeförtolker fremgår af Figur B1, hvor et datasæt med en industriel opdeling af 223 amerikanske virksomheder inden for materiale-sektoren blev uploadet til OpenAI's LLM, ChatGPT.

Figur B2: Eksempel på output genereret vha. ChatGPT's kodeförtolker



Kilde: Eksempel fra ChatGPT's plus-version (tidligere pro), december 2024

Ved hjælp af kodeförtolkeren visualiserede ChatGPT dataene ved at generere Python-koden som vist i Figur B2. Hvis nødvendigt kan denne kode efterfølgende vha. LLM'en omskrives til andre programmeringssprog som STATA, R eller SAS og dermed let integreres i eksisterende arbejde.

Figur B2: Eksempel på Python-kode genereret af ChatGPT's kodeförtolker

```
python Copy code  
  
# Aggregate data by Industry and IndustryShare  
materials_shares = materials_us_data[['Industry', 'IndustryShare']]  
  
# Plotting a pie chart  
plt.figure(figsize=(10, 8))  
plt.pie(  
    materials_shares['IndustryShare'],  
    labels=materials_shares['Industry'],  
    autopct='%1.1f%%',  
    startangle=90  
)  
plt.title('Industry Share in the Materials Sector (US)')  
plt.show()
```

Kilde: Eksempel fra ChatGPT.

## Bilag 2 Eksempel med delopgave fra eksamensopgave

Tabel B1 viser flere af de prompter som ligger imellem den simple prompt og den endelige prompt i Afsnit 3.3. Her skal bemærkes, at den tredje og den femte prompt indeholder samme mængde af information og er bygget op på tilsvarende vis, med den forskel at den ene prompt er på dansk og den anden på engelsk. Selvom de to prompter indeholder den samme information, formår den engelske ChatGPT at oprette et nyt datasæt på korrekt måde, og kommer derudover med en tabeltitel, hvilket ikke er tilfældet med den danske prompt. I tabellen ses kun besvarelsen med datasteps, eftersom den del af besvarelsen, som danner SQL-koden er tilsvarende.

Opgaveformulering: *Træk sognenummeret ud af variabelen sogn i datasættet `sogne_udd_kirke.sas7bdat` i en numerisk variabel kaldet `sognenummer`. Brug både `datasteps` og `SQL`*

Tabel B1: Eksempel på ChatGPT's kodesvar afhængigt af prompt

Prompt	Kodesvar
<p><i>Jeg arbejder i dataprogrammet SAS. Generer kode en kode til følgende opgave: Træk sognenummeret ud af variabelen sogn i datasættet <code>sogne_udd_kirke.sas7bdat</code> i en numerisk variabel kaldet <code>sognenummer</code>. Brug både <code>datasteps</code> og <code>SQL</code>.</i></p>	<pre>data sogne_udd_kirke;     set sogne_udd_kirke;     sognenummer = input(substr(sogn, 1, 4), 8.); /* Ekstraher de første 4 tegn fra 'sogn' og konverter til numerisk */ run;</pre>
<p><i>Jeg er en økonomistuderende, som arbejder i dataprogrammet SAS. Datasættet, jeg arbejder i, hedder <code>sogne_udd_kirke</code> og indeholder oplysninger om befolkningens uddannelsesniveau og bopæl fordelt i sogne. Jeg har lavet et <code>libname</code>, som hedder <code>sasprog</code>. Generer en kode, som trækker sognenummeret ud af variabelen <code>sogn</code> i en numerisk variabel kaldet <code>sognenummer</code>. Jeg vil have koden i både <code>datasteps</code> og <code>SQL</code> og et <code>udprint</code> for at se om koden virker</i></p>	<pre>data sasprog.sogne_udd_kirke;     set sasprog.sogne_udd_kirke;     sognenummer = input(substr(sogn, 1, 4), 4.) /* Ekstraher de første 4 tegn fra 'sogn' og konverter til numerisk */ run; proc print data=sasprog.sogne_udd_kirke (obs=10); /* Udprint de første 10 observationer for at kontrollere */ var sogn sognenummer; run;</pre>
<p><i>Jeg er en økonomistuderende, som arbejder i dataprogrammet SAS. Datasættet, jeg arbejder i hedder <code>sogne_udd_kirke</code> og indeholder oplysninger om befolkningens uddannelsesniveau og bopæl fordelt i sogne. Jeg har lavet et <code>libname</code>, som hedder <code>sasprog</code>, dvs. at jeg forinden har kørt koden: <code>libname sasprog 'path'</code>; Generer en kode til følgende opgave: lav et nyt datasæt, hvor sognenummeret trækkes ud af variabelen <code>sogn</code> og laves til en numerisk variabel kaldet <code>sognenummer</code>. Jeg vil have koden i både <code>datasteps</code> og <code>SQL</code> samt et <code>udprint</code> for at se om koden virker.</i></p>	<pre>data sasprog.nyt_datasaet;     set sasprog.sogne_udd_kirke /* Træk sognenummeret ud af 'sogn' variabelen */     sognenummer = input(scan(sogn, 1, ' '), 8.); run; /* Udprint for at se om koden virker */ proc print data=sasprog.nyt_datasaet (obs=10); run;</pre>
<p><i>Jeg arbejder i dataprogrammet SAS. Datasættet, jeg arbejder i hedder <code>sogne_udd_kirke</code> og indeholder oplysninger om befolkningens uddannelsesniveau og bopæl fordelt i sogne. Jeg har lavet et <code>libname</code>, som hedder <code>sasprog</code>, dvs. at jeg forinden har kørt koden: <code>libname sasprog 'path'</code>; Generer en kode til følgende opgave: lav et nyt datasæt, hvor sognenummeret trækkes ud af</i></p>	<pre>/* DATA step */ data sasprog.sogne_udd_kirke_ny;     set sasprog.sogne_udd_kirke;     /* Antager, at sognenummeret er de første cifre i sogn variabelen */     sognenummer = input(scan(sogn, 1, ' '), 8.); /* Udtræk det første ord (antager det er nummeret) */</pre>

<p>variablen sogn og laves til en numerisk variabel kaldet sognenummer. Jeg vil have koden i både datasteps og SQL samt et udprint for at se om koden virker.</p>	<pre>run; /* Udprint for at tjekke, om det virker */ proc print data=sasprog.sogne_udd_kirke_ny (obs=10);     var sogn sognenummer; run;</pre>
<p><i>I am an economics student working on an assignment in the computer program SAS. The data set I'm working with is called sogne_udd_kirke and contains information regarding the population's educational level in different counties in Denmark. Please generate a SAS code to solve the following: Create a new data set where the county number is extracted from the variable sogn and made into a numerical variable called sognenummer. The county number is the first 4 characters in the variable sogn, and the number and the name of the county are separated by a blank. I would like the code to include both data steps and SQL. Additionally, please include a print statement to check that the code is correct.</i></p>	<pre>libname sasprog 'path'; /* Ensure the library is defined */ data sognenummer_data; /* New dataset name */     set sasprog.sogne_udd_kirke; /* Use the original dataset */     /* Extract the first 4 characters as a numerical variable */     sognenummer = input(substr(sogn, 1, 4), 8.); /* Convert the extracted substring to a number */ run; /* Print the data to verify */ proc print data=sognenummer_data (obs=10); /* Print the first 10 observations */     title "First 10 Observations of the New Dataset sognenummer_data"; run;</pre>

Kilde: Interaktion med ChatGPT's gratisversion, oktober 2024

## De spildte talenter.

Blandt de mange unge danske fodboldtalenter får nogle muligheden for at spille på DBU's udvalgte hold, - men er chancen lige stor for alle?

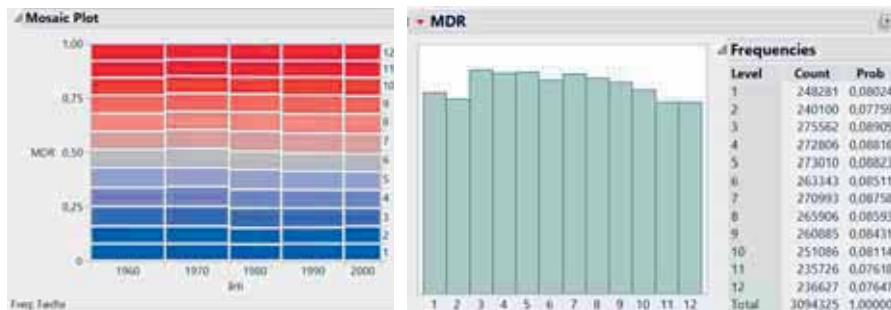
Steen Andersen

Emeritus, Institut for Økonomi, BSS, Aarhus Universitet, e-mail: [sta@econ.au.dk](mailto:sta@econ.au.dk)

**Abstract:** Det kræver et uomtvisteligt talent at blive udvalgt – og talentet udvikler sig over tid. Det betyder, at når man i DBU opererer med årgange, talenter U13-15 og landshold U16-21, så har de, der er født først på året en fordel. I dette papir anvendes data fra DBU's landsholdsdatabase og fokus er på årgangene født mellem 1960 og 2005. Der skelnes ikke mellem hvilke landshold (U16-21) man har spillet på. Der inddrages desuden det aspekt om man har spillet på herrer A-landsholdet. Hvornår man er født året har, (1) større betydning for drenge end for piger, (2) større betydning for årgang 2005 end for årgang 1960 og (3) heldigvis lidt mindre betydning for om man får debut på et A-landshold.

### Fødselsårgangene 1960-2005, fordelt på fødselsmåned.

Nedenstående grafiske illustration indikerer, at fordelingen af fødselsmåned, stort set er uændret i perioden. I begyndelsen af perioden er fødsler i april og maj hyppigere forekomne end i slutningen af perioden. (18,2% i 1960-1979 og 17,0% i 1990-2005). Og begyndelsen af perioden er fødsler i juli og august mindre hyppigt forekomne end i slutningen af perioden. (16,4% i 1960-1969 og 17,6% i 1990-2005).

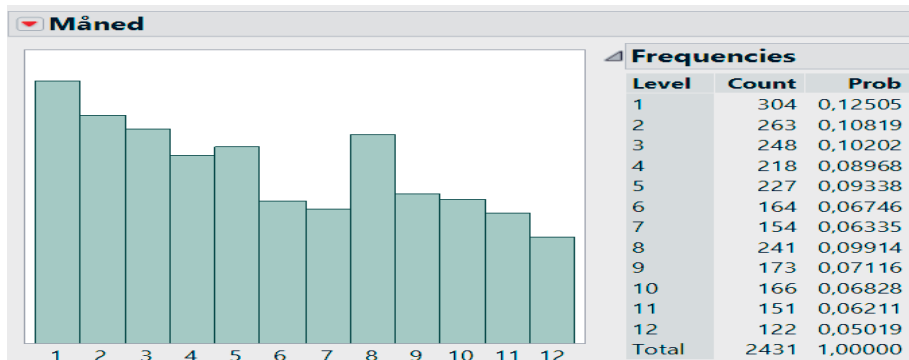


Fordeling af fødte i perioden hen over året er ikke helt jævn, - der er færre fødte i de kolde måneder.



### DBU's udvalgte blandt fødselsårgangene 1960-2005, fordelt på fødselsmåned.

Nedenstående illustrerer tydeligt, at DBU's udvalgte på ingen måde afspejler fordelingen af fødselsmåneder i baggrundsbefolkningen. En der er født i januar har i forhold til en der er født i december 2,4 gange større chance for at være udvalgt ( $304/248281$ )/( $122/236627$ ) til et af DBU's ungdomslandshold.



Der er en markant lavere sandsynlighed for at være udvalgt jo senere på året man er født, - dog afviger de som er født i august ved at have en relativ større chance for at være udvalgt end de, der er født i juni og juli måned og nogenlunde samme chance som de, der er født først på året.

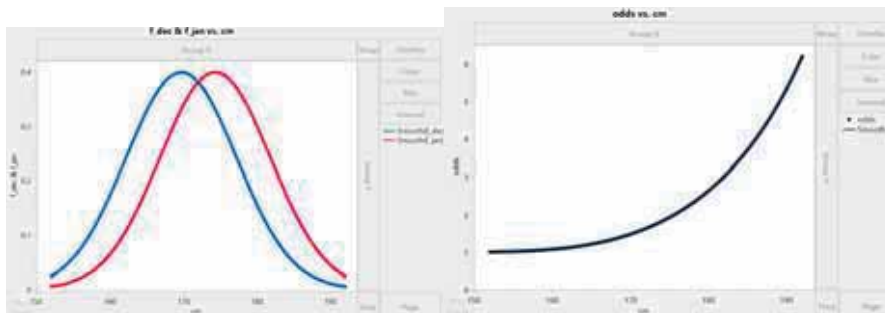
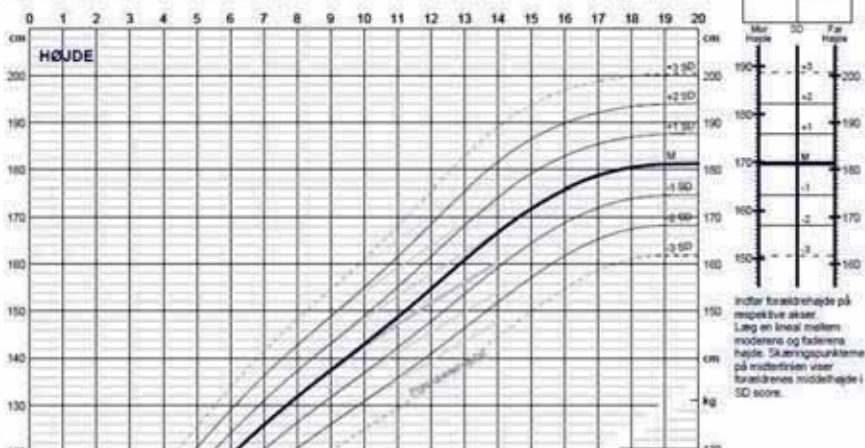
### Vækstkurver for drenge.

Det er ikke fødselsmåned i sig selv, men den fysiske og den mentale fremtoning, der er nogle af mange afgørende parametre for om man er blandt DBU's udvalgte.

Ud fra nedenstående kan man argumentere for, at den forventede højde for 15-årige drenge er 172,0 cm med en standardafvigelse på 7,5 cm. Væksten per år omkring de 15 år er cirka 5 cm, dvs. den forventede højde for U15-drenge født i januar og december er henholdsvis 174,3 cm og 169,7 cm.

CPR nummer	Etternavn	Fornavn	Køn <b>Dreng</b> ♂
------------	-----------	---------	-----------------------

Højde og vægt for alder, fødsel til 20 år



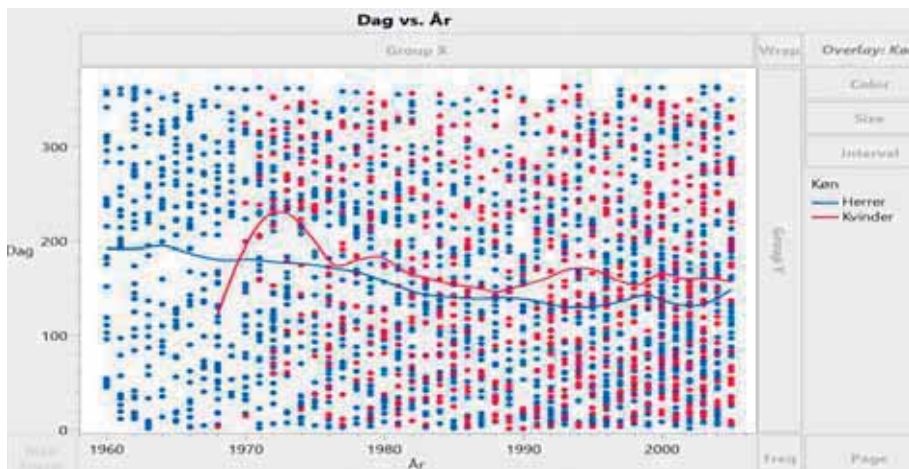
Blandt januar-drengene har 72%, 46% og 22% henholdsvis en højde på over 170, 175, og 180 cm.

Blandt december-drengene har 48%, 24% og 8,5% henholdsvis en højde på over 170, 175 og 180 cm.

Hvilket betyder, at oddset for, at det er en januar-dreng er 1,5, 1,9 og 2,6 givet en højde på henholdsvis 170, 175 og 180 cm – odds som kan sammenholdes med den 2,4 gange større chance nævnt ovenfor – og dermed indikere, at DBU's udvælgelsesprocedure svarer til kun at udvælge blandt spillere, der er 180 cm eller højere!

## DBU's udvalgte fordelt på køn og årgang og den dag på året de er født.

Nedenstående tyder på, at de udvalgte år for år er født lidt tidligere på året end den foregående årgang, mere udpræget for drenge end for piger.



En forklaring herpå kunne være, at der fra 1970 og frem afvikles flere og flere U16/17 kampe, hvor favoriseringen kan antages at være større end ved U18+ kampe. Antallet af kampe varierer meget fra år til grundet afvikling af slutrunder.

Denne sammenhæng kunne man "koge" ned til at;

Response Dag		Parameter Estimates				
<b>Summary of Fit</b>		<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
RSquare	0,032532	Intercept	2922,8563	346,3361	8,44	<.0001*
RSquare Adj	0,031336	Herrer	-20,5154	4,798787	-4,28	<.0001*
Root Mean Square Error	102,7922	År	-1,384538	0,173799	-7,97	<.0001*
Mean of Response	158,3677	(Herrer-0,69601)*(År-1986,65)	-0,314517	0,425818	-0,74	0,4602
Observations (or Sum Wts)	2431					

Dag på året man er født har signifikant betydning i denne sammenhæng - de udvalgte herrer er i gennemsnit født 20 dage tidligere på året end de udvalgte kvinder og dette rykker 1,4 dag per årgang, for både herrer og kvinder.

Gennemsnit for "dag" på året er ved periodens begyndelse tæt på det forventede 183 ((1+365)/2) og derefter signifikant faldende.

Hvis man sammenholder spillere med 1-9 og 10+ ungdomslandskampe ser man

**Response Dag U-Kampe10plus=0**

**Summary of Fit**

RSquare	0.036445
RSquare Adj	0.034268
Root Mean Square Error	103.0179
Mean of Response	160.0736
Observations (or Sum Wgts)	1332

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	533066	177689	16.7430
Error	1328	14093641	10613	<b>Prob &gt; F</b>
C. Total	1331	14626707		<.0001*

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3186.804	441.8384	7.21	<.0001*
Herrer	-21.52432	6.523406	-3.30	0.0010*
År	-1.516263	0.221804	-6.84	<.0001*
Interaction: Herrer*År	0.60744	0.542456	1.12	0.2630

**Response Dag U-Kampe10plus= 1**

**Summary of Fit**

RSquare	0.033979
RSquare Adj	0.031333
Root Mean Square Error	102.3375
Mean of Response	156.3003
Observations (or Sum Wgts)	1099

**Analysis of Variance**

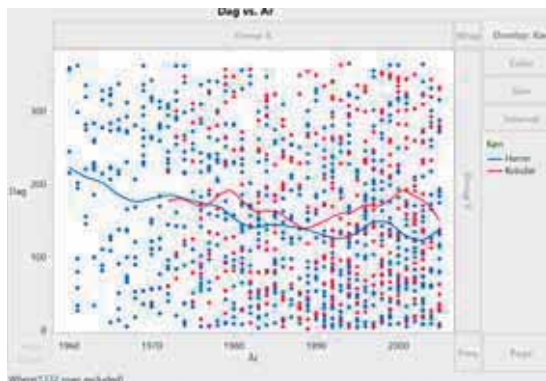
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	403380	134460	12.8388
Error	1095	11467895	10473	<b>Prob &gt; F</b>
C. Total	1098	11871275		<.0001*

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	2410.4364	561.6402	4.29	<.0001*
Herrer	-18.6017	7.077581	-2.63	0.0080*
År	-1.128575	0.281173	-4.01	<.0001*
Interaction: Herrer*År	-1.80573	0.692265	-2.61	0.0092*

- at de etablerede (10+) er født lidt tidligere på året (ikke signifikant), men væsentligst, at der over tid for kvindernes vedkommende ikke sker et ryk for hver årgang mens der for herrerens vedkommende er sket et ryk på  $-1,12 \cdot 1,80 \cdot (1-0,69) = -1,68$  per årgang.

Dette er illustreret lidt mere nuanceret nedenfor:

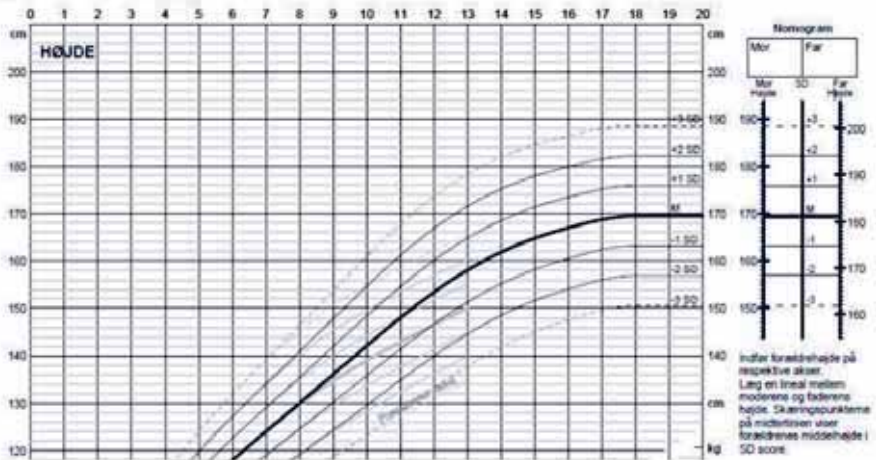


En forklaring herpå kunne være, at scouts/trænere tilknyttet kvindelandsholdene er bedre til at spotte og udvikle talenter, at der er langt færre piger end drenge<sup>\*)</sup>, der spiller fodbold eller det faktum, at pigers vækstkurve flader ud tidligere end drenges.

<sup>\*)</sup> I 2023 var 62% af DBU's medlemmer under 19 år og 80 % af DBU's medlemmer var drenge/mænd.

CPR nummer	Efternavn	Fornavn	Køn <b>Pige</b> ♀
------------	-----------	---------	----------------------

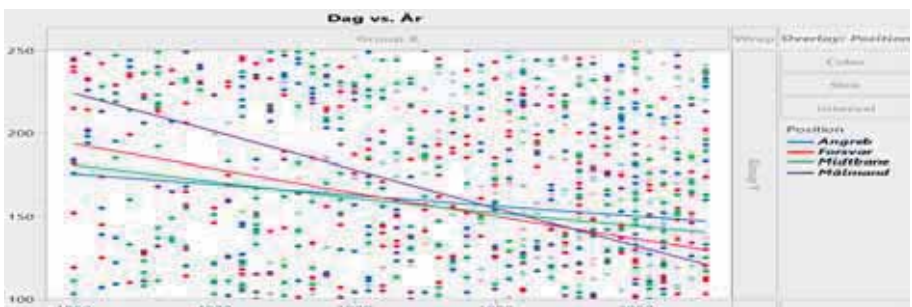
### Højde og vægt for alder, fødsel til 20 år



DBU er opmærksom på problematikken og har initiativer, der understøtter såkaldte ”late developers”.

### Har positionen på banen nogen betydning?

For herrer, vist nedenfor, har dag på året større betydning over tid jo tættere man er på målet (p-værdi ved test er 0,027). Samlet set for herrer reduceres dag født på året med 1,3. For målmænd er det med 2,3, for forsvarer 1,4, for midtbane 0,9 og for angribere 0,6.



## Hvem er det så, der får debut på A-landsholdet?

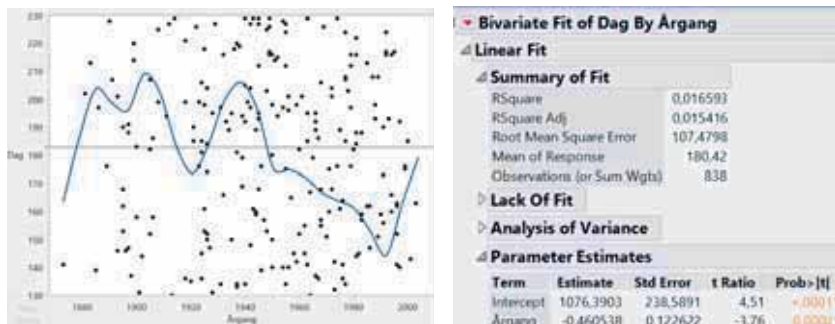
Fra DBU's spillerdatabase uddrages de 838 spillere der til dato (december 2024) har haft debut på herrer A-landsholdet.

Der er set bort fra de 17 spillere der udgjorde vikar-landsholdet samt 4 kvinder, der fejlagtigt er blevet rubriceret under herrer A-landsholdet.

Fra første A-landskamp og frem til seneste (nr. 914) har i alt 838 spillere optrådt på A-landsholdet.

DATO	RESULTAT	MODSTANDER	DATO	RESULTAT	MODSTANDER
19-10-1908	9 - 0	Frankrig	18-11-2024	0 - 0	Serbien

For disse spillere ser man også, som for ungdomsspillerene, at de over tid er født tidligere på året:

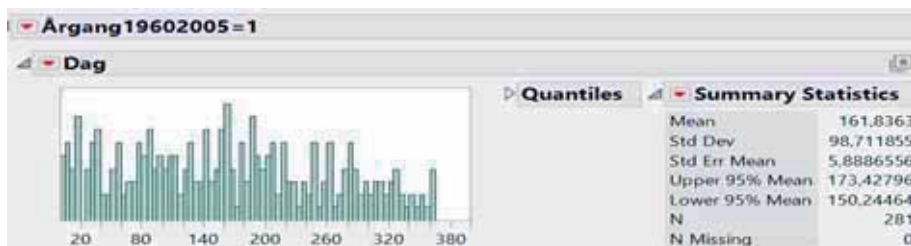


Det er dog ikke så markant som for herrer ungdomsspillerne årgang 1960-2005. De var i gennemsnit født dag 154 og faldet per år var 1,5 dag. For de samme årgange, der A-debutere var tilsvarende tal dag 162 og med et insignifikant fald på 0,6 dag per år:

Summary of Fit		Parameter Estimates				
RSquare	0,005594	Term	Estimate	Std Error	t Ratio	Prob> t
RSquare Adj	0,002029	Intercept	1371,1029	965,3078	1,42	0,1566
Root Mean Square Error	98,61164	Årgang	-0,610971	0,487704	-1,25	0,2113
Mean of Response	161,8363					
Observations (or Sum Wgts)	281					



Det skal dog pointeres, at dag 162 er signifikant lavere end baggrunds populationens dag 183:



Forskellen på 162 og 183 svarer til at man havde udvalgt helt ligeligt blandt talenter født mellem 1. januar og 20. november og havde set helt bort fra dem, der var født mellem 21. november og 31. december – det er omkring 10% af en årgang!

Opdelt på hvor etablere spillerne er (få, nogle eller mange A-landskampe) er der ingen yderligere fund. Med hensyn til spillernes position på banen er der en tendens (men ikke signifikant) til, at målmænd født tidligt på året favoriseres mere end de på andre positioner.

### Konklusion.

Det er primært for de yngste ungdomsherrer (og specielt målmand/forsvar) at de, der er født tidligt på året, favoriseres. Når man bevæger sig fra U16 til U21-landshold bliver den relative forskel i højde (vækst) på om man er født tidligt eller senere på året mindre og mindre. De tidlige fravalg på udvalgte hold af talenter, der er født sent på året, har konsekvens for det fremtidige karriereforsøg.

For herrer A-landsholdsspillere svarer denne favorisering til, at man ser bort fra 10 % af en årgang.

Talentet kan være dog være meget åbenlyst, spilleren kan være vedholdende og for nogens vedkommende modnes/hærdes man måske af de udfordringer, man som "late developper" stilles overfor i teenageårene.

Som tidligere nævnt er fødselsdatoen ikke eneste indikator, for om man er "late developper". Fokus her, har udelukkende været at se på spillere på udvalgte hold i DBU-regi. Favoriseringen af "tidligt fødte", der spiller på lavere niveau, er ikke berørt her. Man kunne forstille sig, at der hvor talentet ikke rækker eller der hvor opmærksomheden fra træner-side ikke er til stede, - der løber de "sent fødte" hurtigt panden imod muren og lader på et tidspunkt støvlerne blive på hylden.

# Hvem har gode matematikkskaber trods matematikskræk

Anders Milhøj

Anders.milhøj@econ.ku.dk

I PISA undersøgelsen spørges til elevens egen holdning til matematikkskaber med spørgsmålet: *ST268Q04JA = Agree/disagree: Mathematics is easy for me*. Svarene gives på en skala fra 1 til 4:

1. Strongly disagree
2. Disagree
3. Agree
4. Strongly agree

Dette spørgsmål er ikke besvaret af 1071 elever.

Vi sammenholder svarene på dette spørgsmål med den gennemsnitlige matematikscore, *score\_m*, der er gennemsnittet af de 10 plausible values, *pv1math - pv10math*. Alle er tildelt en matematikscore. Gennemsnittet af matematikscoren i Danmark er 478.8. Ud fra dette tal inddeles eleverne i to grupper, 'høj' og 'lav', alt efter om scoren er over eller under dette gennemsnit. Bemærk, at data er indsamlet stratificeret efter indvandrerbaggrund med overrepræsentation af skoler med mange ikke-etnisk danske elever. Dette giver en vis overrepræsentation af bogligt svage elever, så gennemsnittet 478.8 er noget lavere end det sande gennemsnit af danske skoleelevers matematikscore.

Tabellen viser fordelingen af svarene på dette spørgsmål sammenholdt med matematikscoren kategoriseret i høj/lav. De nederste tal i cellerne er rækkeprocenter.

De fleste elever, der vurderer matematik som let (Agree eller Strongly agree), har en høj matematikscore. Omvendt har en stor del af de elever, der finder matematik svært (Disagree eller Strongly Disagree), en lav matematikscore, hvilket jo ikke kan for-bavse. Dette indikerer en sammenhæng mellem elevernes selvopfattelse og deres faktiske præstationer i matematik.

Der findes alligevel 166 atypiske elever, der har svaret 'Strongly disagree' på spørgsmålet om egne matematikevner, men som alligevel har en matematikscore over gennemsnittet, mens de 600 andre elever, der har svaret 'Strongly disagree', har den mere forventede matematikscore under gennemsnittet. Disse 166 elever betegnes i det følgende som 'Atypiske', mens de 600. for hvem matematikevnerne svarer bedre til deres egen fornemmelse af lave matematikkskaber, betegnes 'Normale'.



<i>Mathematics is easy for me.</i>	<i>Høj</i>	<i>Lav</i>	<i>Total</i>
<i>Strongly disagree</i>	166 21.67	600 78.33	766
<i>Disagree</i>	658 38.37	1057 61.63	1715
<i>Agree</i>	1307 66.31	664 33.69	1971
<i>Strongly agree</i>	558 82.42	119 17.58	677
<i>Total</i>	2689	2440	5129
<i>Frequency Missing = 1071</i>			

Blandt de 766 elever med matematiksræk er kønsfordelingen meget skæv med 68% piger, men kønsfordelingen mellem atypiske elever og normale elever er helt ens,  $p = 0.95$ .

I det følgende sammenlignes de to gruppers svar på andre relevante spørgsmål i datasættet. Disse andre spørgsmål er først og fremmest udvalgt ved at screene for deres evne til at adskille de to elevgrupper med resp. høj og lav matematikscore blandt de elever, der har svaret 'Strongly disagree' på spørgsmålet om egne matematikevner, ved hjælp af en SAS-makro. Listen over spørgsmål er dog på ingen måde komplet, fx er der kun valgt et enkelt spørgsmål af mange stort set ensartede spørgsmål. I de viste resultater er antal observationer forskellige, da der er forskelligt bortfald på de stillede spørgsmål.

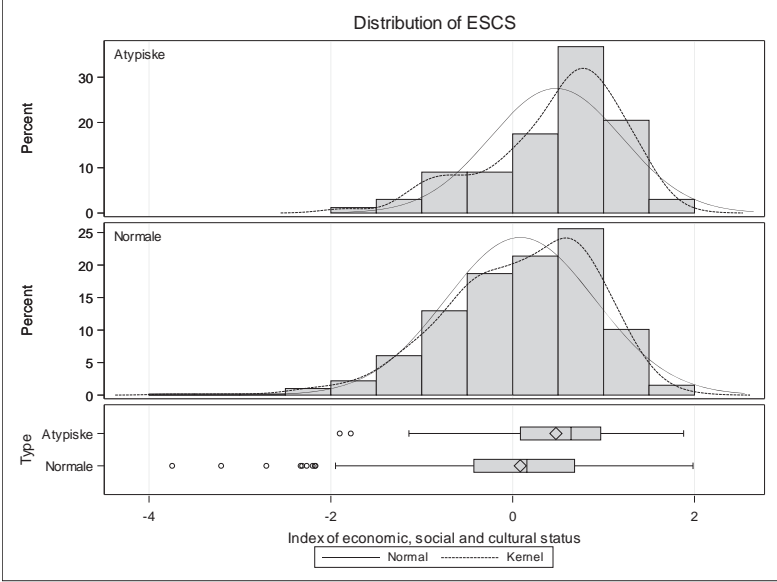
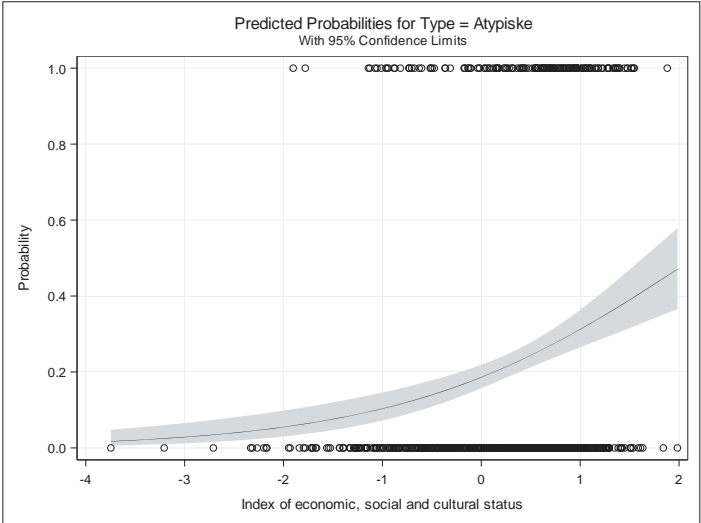
### **Socioøkonomiske forklaringer**

I datasættet findes en indikator for hjemmets resurser, ESCS, som er sammensat af oplysninger om forældrenes uddannelse, indkomst samt elevens angivelser af akademiske/kulturelle ting i hjemmet, fx. musikinstrumenter og bøger. Dette indeks er konstrueret som en standardiseret normalfordelt variabel. I en logistisk regressionsanalyse ses, at denne variabel har stor betydning for sandsynligheden for at opnå atypisk høj matematikscore trods skrækken for matematik.

Figuren viser at denne sandsynlighed vokser markant til ca. 50% for eleverne med højest værdi af ESCS. Forældre med overskud burde altså bruge noget af deres overskud på at bearbejde deres børns matematiksræk.

Den anden figur illustrerer samme sammenhæng ud fra en anden tankegang. Selvom en eventuel kausalitet nok ikke kan gå fra matematikscoren til matematiksræk, er det

interessant at se, hvordan ESCS er forskellig for de atypiske elever og de normale elever. Gennemsnittet af ESCS er 0.08 for normale elever og 0.48 for de atypiske elever – og denne forskel er klart signifikant  $t = 5.6$ . Figuren sammenholder de to histogrammer for de atypiske hhv. normale elever.



## Indvandrerbaggrund

Immigrant baggrunden er angivet på tre trin efter en OECD definition.

<i>Immigrant background</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Native student</i>	144 24.08	454 75.92	598
<i>Second-Generation student</i>	13 11.40	101 88.60	114
<i>First-Generation student</i>	7 17.50	33 82.50	40
<i>Total</i>	164	588	752

*Frequency Missing = 14*

De fleste atypiske elever er etnisk danske, mens normale elever har en højere andel af første og anden generations elever. Denne tabel er relateret til den næste tabel om sprog.

## Det talte sprog i hjemmet

<i>Language at home</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Language of the test</i>	151 23.78	484 76.22	635
<i>Other language</i>	15 11.54	115 88.46	130
<i>Total</i>	166	599	765

*Frequency Missing = 1*

En dobbelt så stor andel af de elever, der trods matematikskræk, opnår en høj matematikscore blandt de elever, der dansk i hjemmet, end blandt andre elever. I den forbindelse skal man huske, at et andet sprog jo godt kan være fx engelsk, så et andet sprog ikke nødvendigvis er et tegn på et ikke-akademisk hjem. Tallene kunne dog godt tyde på, at en del af matematikopgaverne er så verbalt formulerede, at elever, der ikke er så gode til dansk, har et handicap i matematikopgaverne.

## Pjæk

<i>I skipped a whole school day the last two weeks</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Never</i>	134 25.43	393 74.57	527
<i>One or two times</i>	26 16.56	131 83.44	157
<i>Three or four times</i>	2 6.45	29 93.55	31
<i>Five or more times</i>	4 10.26	35 89.74	39
<i>Total</i>	166	588	754

*Frequency Missing = 12*

Atypiske elever pjækker mindre end normale elever. Nu er det måske en overfortolkning at opfatte dette spørgsmål, som et spørgsmål om pjæk, for fravær kunne vist efter spørgsmålets formulering også være sygdom. Men det er dog påfaldende, at elever helt uden fravær har en højere sandsynlighed for et godt matematikresultat end elever med fravær.

## Elevers mening om lærerens arbejde

<i>My teachers are interested in my wellbeing</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Strongly disagree</i>	1 2.86	34 97.14	35
<i>Disagree</i>	13 14.29	78 85.71	91
<i>Agree</i>	57 22.35	198 77.65	255
<i>Strongly agree</i>	25 33.33	50 66.67	75
<i>Total</i>	96	360	456

*Frequency Missing = 310*

Atypiske elever føler i højere grad, at deres lærere er interesserede i deres velbefindende sammenlignet med normale elever. Det har altså en positiv indflydelse på deres resultater, at de føler, at læreren oprigtigt interesserer sig for deres velbefindende. Det er også glædeligt, at det ikke er så mange, der føler, at lærerne ikke interesserer sig for deres velbefindende.

### Udenadslæren er en god ting!

<i>Memorize mathematical rules</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Never or almost never</i>	12 21.82	43 78.18	55
<i>Less than half of the lessons</i>	12 16.67	60 83.33	72
<i>About half of the lessons</i>	21 19.63	86 80.37	107
<i>More than half of the lessons</i>	24 24.00	76 76.00	100
<i>Every lesson or almost every lesson</i>	26 35.62	47 64.38	73
<i>Total</i>	95	312	407
<i>Frequency Missing = 359</i>			

Hvis der indgår elementer af 'klassisk' udenadslæren i matematikundervisningen, klarer eleverne sig bedre. Hvad der menes med *mathematical rules* i niende klasse, er selvfølgelig uklart, for det kan vel være alt fra den lille tabel til formelen for en cirkels areal, så det skal vel fortolkes som, at et fast ståsted øger mulighederne for at løse opgaver.

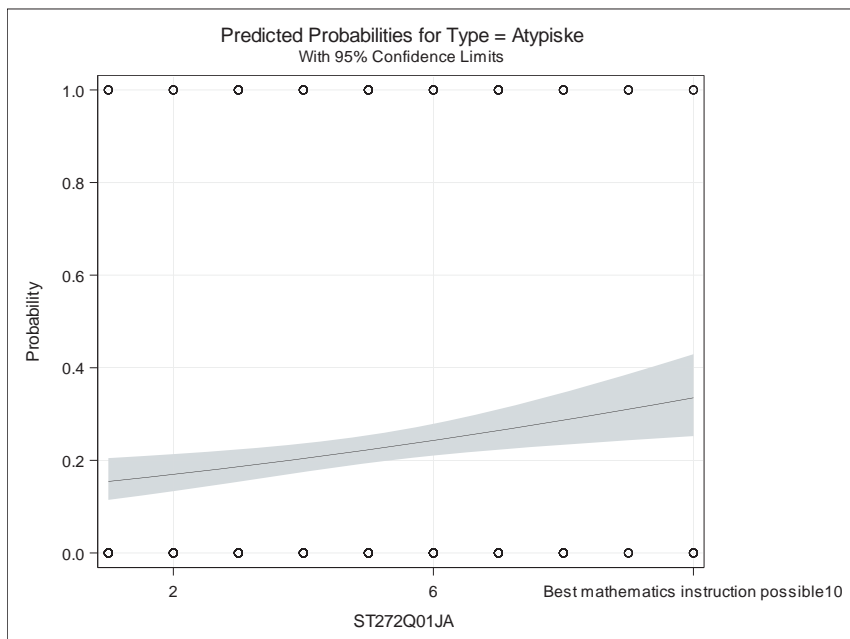
### Løse ligninger og andre aspekter af matematikundervisningen

Ligningen i spørgsmålet,  $2(x+3)=(x+3)(x-3)$ , er i øvrigt lidt drilsk, så det er ret oplagt at løse den forkert ved at overse løsningen  $x = -3$ . Men det er oplagt, at hvis undervisningen udfordrer eleverne, løftes niveauet. Det sker heldigvis i mere end halvdelen af skolerne, hvis man skal bedømme det efter elevernes svar, hvad der dog kan være tvivlsomt

<i>Solving an equation like <math>2(x+3)=(x+3)(x-3)</math></i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Frequently</i>	53 25.48	155 74.52	208
<i>Sometimes</i>	28 21.05	105 78.95	133
<i>Rarely</i>	5 17.24	24 82.76	29
<i>Never</i>	1 4.35	22 95.65	23
<i>Total</i>	87	306	393

*Frequency Missing = 373*

Et overordnet spørgsmål om *Quality of mathematics instruction*, hvor der svares på en skala fra et til ti, giver en signifikant sammenhæng med en odds ratio på 1.119. Sammenhængen illustreres ved følgende effectplot, der viser, at hvis eleverne synes matematikundervisningen er *best possible* er sandsynligheden for at få en høj matematikscoren i PISA undersøgelsen ca 30%, mens den kun er knapt 20% for elever, der mener undervisningen er ringe.



## Computer i matematiktimer

<i>How often using computer mathematics systems in school</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>Frequently</i>	56 32.00	119 68.00	175
<i>Sometimes</i>	24 16.55	121 83.45	145
<i>Rarely</i>	8 20.51	31 79.49	39
<i>Never</i>	2 9.52	19 90.48	21
<i>Total</i>	90	290	380

*Frequency Missing = 386*

Eleverne motiveres åbenbart af brug af IT-baserede værktøjer i matematikundervisningen, så heldigvis sker det på de fleste skoler i datasættet. Det er også påfaldende, at elever, der ofte bruger computer til matematik, klarer sig bedre, end elever, der ikke gør det, så de fordele ved udenadslæren, der blev beskrevet tidligere, bliver ikke hæmmet ved brug af IT-værktøjer.

Men svarene på andre spørgsmål viser, at brug af mobiltelefoner til at følge med på sociale medier i timerne er skadeligt. Det er altså ikke godt at være 'på' i mere end halvdelen af timerne, for det mindsker sandsynligheden for at opnå en god matematikscore. De atypiske elever er mere enige i, at der skal sættes regler for privat brug af mobiltelefoner i timerne, end de normale elever med matematikkræk.

En anden ramme for spørgsmål viser, at elever, der anvender IT-værktøjer ved lektielæsning hjemme, klarer sig bedre end elever, der sjældent gør det. Svarene på en række spørgsmål viser også, at elever der mere end to timer dagligt, er på sociale medier eller bruger skærme til underholdning, fx spil eller film, klarer sig dårligere. Men dette resultat er bestemt ikke entydigt, hvad følgende tabel viser.

<i>Facebook®, Instagram®, Twitter®, emails, chat</i>	<i>Atypiske</i>	<i>Normale</i>	<i>Total</i>
<i>No time at all</i>	11 25.58	32 74.42	43
<i>Less than 1 hour a day</i>	24 21.43	88 78.57	112
<i>Between 1 and 3 hours a day</i>	56 37.09	95 62.91	151
<i>More than 3 hours and up to 5 hours a day</i>	36 21.05	135 78.95	171
<i>More than 5 hours and up to 7 hours a day</i>	17 15.04	96 84.96	113
<i>More than 7 hours a day</i>	7 12.96	47 87.04	54
<i>Total</i>	151	493	644

*Frequency Missing = 122*

### **Ihærdighed og glæde ved skolen**

<i>I stop when work becomes too difficult</i>	<i>Høj_disagree</i>	<i>Normal</i>	<i>Total</i>
<i>Strongly disagree</i>	6 35.29	11 64.71	17
<i>Disagree</i>	33 42.86	44 57.14	77
<i>Neither agree nor disagree</i>	29 20.14	115 79.86	144
<i>Agree</i>	17 15.32	94 84.68	111
<i>Strongly agree</i>	3 11.11	24 88.89	27
<i>Total</i>	88	288	376

*Frequency Missing = 390*



De ret få elever, der hurtigt giver op, klarer sig dårligt, og et yderligere spørgsmål om at fuldføre opgaver, der viser sig komplicerede, viser samme klare tendens. Glæden ved at lykkes med noget er altså væsentlig for lysten til at lære mere og bedre, mens gentagne fiaskoer giver eleverne lyst til at opgive det hele. Et yderligere spørgsmål viser, at elever, der har lysten til at lære noget nyt, lærer sig bedre, hvilket nok kan tilskrives samme tendens, at faglig mistroivsel ødelægger indlæringen.

## **Konklusion**

VIVE udgav i november 2024 en rapport, Jóelsdóttir, L.B. & Østergaard, K.: *PISA 2022 – Delrapport matematik. Danske 15-åriges oplevelse af deres matematikundervisning – og deres motivation, selvsikkerhed, matematikangst og mindset*. Den omhandler lignende emner og sammenligner bla. med andre OECD lande med tilsvarende konklusioner ved brug af gennemsnit og andele blandt hele populationen af alle elever.

I dette symposieindlæg er der udelukkende fokus på elever, der trods matematikkræk alligevel klarer sig godt i PISAs matematiske spørgsmål. Mange af de viste sammenhænge er oplagte, og de er som oftest netop sammenhænge og ikke kausaliteter. Da de gode matematikevner hos en elev med matematikkræk kan være et tegn på, at eleven trods alt er en god, ihærdig og dygtig elev, der bare fejlagtigt føler ikke helt at kunne slå til i matematik. Overvægten af piger i gruppen med matematikkræk kunne tyde på, at det kunne være tilfældet, så det ville være en god idé at undersøge om effekterne i de mange tabeller vekselvirker med kønnet.

## **Aktivitsdeltagelse, dødelighed og hospitalsindlæggelser: Et dansk kohortestudie**

Linda Juel Ahrenfeldt<sup>1</sup>, Tobias Anker Stripp<sup>1,2,3</sup>, Jens Søndergaard<sup>1</sup>, og Søren Möller<sup>4,5</sup>

- 1) Forskningsenheden for Almen Praksis, Institut for Sundhedstjenesteforskning, Syddansk Universitet, Esbjerg-Odense, Danmark
- 2) The Human Flourishing Program, Quantitative Institute for Social Sciences, Harvard Universitet, Cambridge, MA, USA
- 3) Center for Videnskab og Tro, Københavns Universitet, København, Danmark
- 4) Open Patient data Explorative Network, Odense Universitetshospital, Odense, Danmark
- 5) The OPEN Research Unit, Klinisk Institut, Syddansk Universitet, Odense, Danmark

Baseret på artiklen *Impact of activity participation on the risk of mortality and hospitalisations in Danish men and women: Insights from REGLINK-SHAREDK* <sup>1</sup> accepteret i *Epidemiology* i oktober 2024.

### **Introduction**

Med længere levetid og en voksende ældrebefolkning er det vigtigt at fokusere på at forbedre sundheden blandt de ældre. Banebrydende studier af Berkman og Syme i 1979 <sup>2</sup> og House et al. i 1988 <sup>3</sup> viste den beskyttende effekt af sociale relationer på dødelighedsrisikoen. Efterfølgende epidemiologiske undersøgelser, herunder systematiske oversigtsartikler og meta-analyser, har konsekvent bekræftet sammenhængen mellem social tilknytning, sundhed og dødelighed <sup>4-6</sup>.

Mens social tilknytning omfatter forskellige former for engagement i det sociale miljø, har en specifik forskningsgren undersøgt formel social deltagelse, som kendetegnes ved aktiv involvering i etablerede organisationer <sup>7</sup>. Det har dog vist sig, at formel social deltagelse har forskellige sundhedseffekter afhængigt af typen af aktivitet. For eksempel er frivilligt arbejde forbundet med lavere dødelighed <sup>8</sup>. Omvendt er omsorg eller pleje af familiemedlemmer ofte stressende og kan påvirke både fysisk og mental sundhed negativt <sup>9</sup>, og nogle studier tyder på at omsorg og pleje hænger sammen

med højere dødelighed <sup>10</sup>. Forskning, der undersøger sammenhængen mellem aktivitetsdeltagelse og brug af sundhedsydelser, såsom indlæggelser, er stadig begrænset.

For nyligt undersøgte vi sammenhængen mellem deltagelse i en religiøs organisation samt dødelighed og indlæggelser blandt 2987 danskere på 40 år og derover, som blev interviewet fra 2004-2007 og fulgt i de danske registre frem til 2018 <sup>11</sup>. Vi observerede, at personer, der havde deltaget i religiøse organisationer inden for den seneste måned, havde en 30 % lavere dødelighed, med en mere markant sammenhæng blandt kvinder (44 %), men ingen signifikant sammenhæng blandt mænd. Omvendt havde mænd, der deltog i en religiøs organisation, 33 % færre indlæggelser per år, mens en lignende sammenhæng ikke blev observeret blandt kvinder <sup>11</sup>. Selvom undersøgelsen bidrog til den begrænsede forskning om de potentielle sundhedsmæssige fordele ved deltagelse i religiøse organisationer, står ét spørgsmål tilbage: Skyldes disse sundhedseffekter det at være religiøst aktiv, eller kan vi finde lignende sammenhænge for andre aktiviteter?

I dette studie undersøger vi sammenhængen mellem aktivitetsdeltagelse og dødelighed og hospitalsindlæggelser blandt midaldrende og ældre danskere. Vores hypotese er, at formel social deltagelse er forbundet med lavere dødelighed og færre indlæggelser, med de stærkeste sammenhænge for dødelighed blandt kvinder og indlæggelser blandt mænd.

## **Metoder**

**Datakilde og studiepopulation** The Survey of Health, Ageing and Retirement in Europe (SHARE) er en omfattende survey-undersøgelse, der indsamler information om den aldrende befolkning i Europa gennem personlige interviews <sup>12</sup>. SHARE indsamler bl.a. data om sundhed, økonomisk status og sociale faktorer for personer på 50 år og derover samt deres ægtefæller eller partnere uanset alder. Undersøgelsen startede i 2004-2005 og er siden da gennemført hvert andet år <sup>12</sup>. I REGLINK-SHAREDK-projektet er danske SHARE-data blevet koblet med danske registerdata på Danmarks Statistik, hvilket skaber et unikt datasæt, der kombinerer subjektive interviewdata med objektive registerdata <sup>13</sup>.

I denne longitudinelle undersøgelse inkluderede vi danske mænd og kvinder på 40 år eller ældre, som havde deltaget i mindst én SHARE-interviewrunde i enten 2004-2005 eller 2006-2007. Vi ekskluderede 2 personer under 40 år, 54 personer uden svar på spørgsmål om aktiviteter, 29 personer hvis survey-data ikke kunne kobles til de danske registre, samt én person uden interviewdato, hvilket resulterede i en studiepopulation på 2987 personer, identisk med studiepopulationen i vores tidligere artikel <sup>11</sup>.

**Dødelighed og indlæggelser** De danske SHARE-data blev koblet til nationale sundhedsregistre ved hjælp af CPR-numre. Vi anvendte tre registre: CPR-registret, Dødsårsagsregisteret og Landspatientregisteret.

**Aktivitetsdeltagelse** Data om aktivitetsdeltagelse kom fra SHARE-interviews, hvor deltagerne blev spurgt om deres deltagelse i syv forskellige aktiviteter indenfor den sidste måned. Deltagerne blev også spurgt om hyppigheden af disse aktiviteter og deres årsager til deltagelse.

**Statistiske analyser** Vi anvendte generaliserede lineære modeller med pseudo-observationsbaserede estimater til at evaluere dødelighedsforskelle ved 90 års alderen for personer der havde deltaget i aktiviteter indenfor den seneste måned ift. dem der ikke havde <sup>14,15</sup>. Denne metode anvender en cloglog-regressions-model sammen med Kaplan-Meier-estimatoren  $S(t_l)$  på valgte tidspunkter  $t_l$  (i vores tilfælde  $t_1 = 90$  år) til at estimere relative forskelle:

$$cloglog(S(t_l)) = \log(-\log(S(t_l))) = \beta^T Z_l$$

og en logistisk regressionsmodel sammen med Aalen-Johansen estimatoren  $Q(t_l)$  på tidspunktet  $t_l$ :

$$logit(Q(t_l)) = \beta^T Z_l$$

til at estimere absolutte forskelle, baseret på de modellerede sandsynligheder.

Baseret på disse modeller estimerede vi relative risici (RR) og risikodifferencer (RD) med 95% konfidensintervaller (CI). Hospitaliseringsmålet omfattede alle heldøgnslæggelser registreret i løbet af den 14-årige opfølgingsperiode. På grund af overdispersion i Poisson-regressionsmodellen anvendte vi negativ binomial regression

til at beregne incidensrate-ratioer (IRR) for indlæggelser. Studieperioden begyndte på interviewdatoen og sluttede enten den 31. december 2018 eller ved deltagernes død, afhængigt af hvad der indtraf først.

Vi estimerede to modeller. Model 1 justerede for alder, køn og interviewbølge. Model 2 justerede også for uddannelse, civilstand, formue, medicinsk historie, rygning, alkoholforbrug, BMI, depressive symptomer og begrænsninger i dagligdags aktiviteter (ADL). I alle analyser vurderede vi effekterne separat for hvert køn ved at inkludere en interaktionsterm mellem aktivitetsdeltagelse og køn.

For at vurdere robustheden af resultaterne lavede vi fire sensitivitetsanalyser baseret på model 2.

1) Beregning af E-værdier ift. resultaterne fra hovedanalyserne. Denne analyse blev udført ved brug af formlen

$$E - value = HR + \sqrt{HR \cdot (HR - 1)}$$

Baseret på den oprindelige artikel af VanderWeele og Ding<sup>16</sup> ved hjælp af Stata-pakken "evalue"<sup>17</sup>.

2) Ekskludering af deltagere der havde deltaget i en religiøs organisation.

3) Ekskludering af deltagere med ADL-begrænsninger ved baseline.

4) Hospitaliseringsmålet udvides til også at omfatte dagpatienter, ambulante patienter og skadestuebesøg.

Alle analyser blev udført ved hjælp af Stata version 17.0.

## Resultater

Studiepopulationen bestod af 1.348 mænd (45,1%) og 1.639 kvinder (54,9%) med en gennemsnitsalder på 62,6 år. I alt deltog 36,4% i en sports- eller socialklub, 30,8% hjalp familie og venner, 19,4% udførte frivilligt arbejde, 10,6% deltog i et uddannelses- eller træningskursus, 5,9% deltog i en religiøs organisation, 5,8% passede en syg eller handicappet person, og 4,9% var involveret i en politisk eller samfundsrelateret organisation (Tabel 1).

Tabel 1 - Baselinekarakteristika for studiepopulationen.

	Total	Mænd	Kvinder
Antal (%)	2.987	1.348 (45,1)	1.639 (54,9)

Alder (år), gennemsnit (yngst-ældst)	62,6 (40-98)	62,5 (43-94)	62,8 (40-98)
Døde	848 (28,4)	407 (30,2)	441 (26,9)
Deltagelse i aktiviteter			
Frivilligt/velgørende arbejde	578 (19,4)	285 (21,1)	293 (17,9)
Passet syg/handicappet	172 (5,8)	63 (4,7)	109 (6,7)
Hjulpet familie/venner	919 (30,8)	454 (33,7)	465 (28,4)
Uddannelses/træningskursus	315 (10,6)	121 (9,0)	194 (11,8)
Sports/socialklub	1.083 (36,4)	486 (36,1)	597 (36,4)
Politisk/samfundsrelateret organisation	147 (4,9)	87 (6,5)	60 (3,7)
Religiøs organisation	176 (5,9)	68 (5,0)	108 (6,6)
Samlet aktivitetscore			
0 aktiviteter	1014 (34,0)	441 (32,7)	573 (35,0)
1 aktivitet	1042 (34,9)	472 (35,0)	570 (34,8)
2 aktiviteter	587 (19,7)	275 (20,4)	312 (19,0)
3+ aktiviteter	344 (11,5)	160 (11,9)	184 (11,2)

Tabel 2 skitserer de vigtigste motivationsfaktorer for de forskellige aktiviteter. Den vigtigste grund til at deltage i frivilligt arbejde var at yde noget nyttigt (72,7%). Den primære motivation for at passe en syg eller hjælpe familie og venner var at føle sig nødvendig (henholdsvis 87,8% og 83,2%). Uddannelses- eller træningsaktiviteter var drevet af ønsket om at bruge færdigheder eller holde sig i form (60,0%). Ift. sports- eller sociale klubber var fornøjelse (78,8%) og at holde sig i form (75,4%) de vigtigste drivkræfter. Politisk engagement var primært drevet af ønsket om at bidrage (64,0%), mens deltagelse i religiøse organisationer hovedsageligt var for personlig fornøjelse (65,9%) (Tabel 2).

*Tabel 2 - Årsager til aktivitetsdeltagelse angivet i procent.*

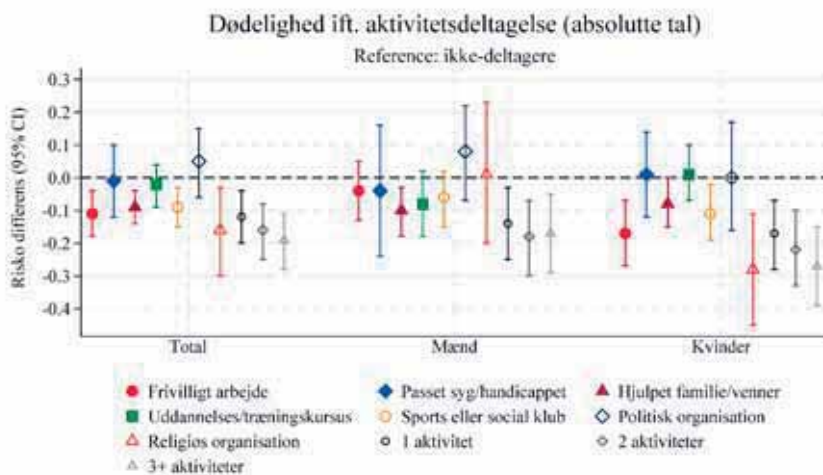
	Frivilligt arbejde	Passet syg/handicappet	Hjulpet familie/venner	Uddannelses/træningskursus	Sports/socialklub	Politisk organisation	Religiøs organisation
For at møde folk	50,5	2,9	11,3	19,4	55,8	44,2	46,9
For at yde noget nyttigt	72,7	32,0	40,0	15,9	9,8	64,0	28,0
For at opnå noget personligt	23,7	4,4	5,3	50,6	26,8	20,5	49,4
Fordi der er brug for mig	63,7	87,8	83,2	15,2	9,7	37,4	27,4
For at tjene penge	1,6	5,2	0,3	9,2	0,4	2,0	-
Fordi jeg nyder det	68,3	22,0	46,6	34,5	78,8	50,6	65,9
For at bruge mine evner eller holde mig i form	41,7	7,6	15,8	60,0	75,4	29,3	16,6
Fordi jeg føler, jeg er tvunget til at gøre det	5,0	14,3	5,0	18,4	3,0	9,6	-

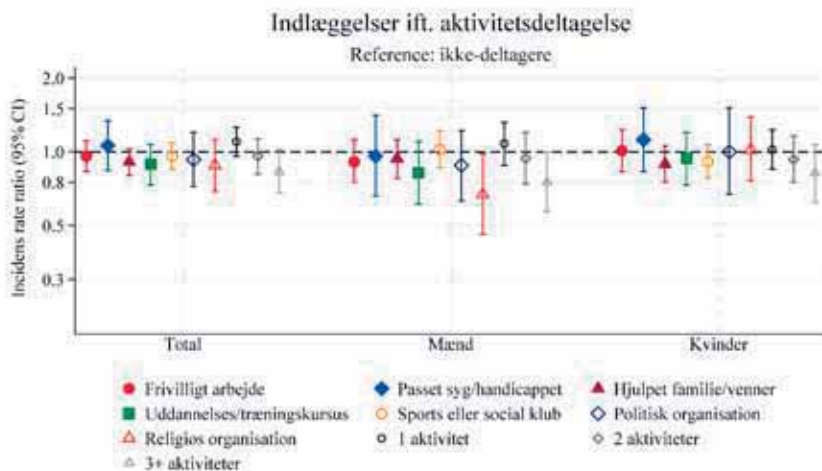
-Tallene kunne ikke estimeres pga. mindre end fem personer i grupperne.

Blandt de 848 registrerede dødsfald viste model 2 lavere dødelighed for personer, der deltog i 1, 2 eller 3+ aktiviteter sammenlignet med dem, der ikke deltog. Laveste

dødelighed var indikeret blandt dem, der deltog i 3+ aktiviteter (RR 0,78; 95% CI 0,67–0,89; RD -0.19; 95% CI -0.28 to -0.11). Specifikke aktiviteter, der var associeret med lavere dødelighed, inkluderede frivilligt arbejde (RR 0,85; 95% CI 0,76–0,95; RD -0.11; 95% CI -0.18 to -0.04), hjælp til familie og venner (RR 0,88; 95% CI 0,82–0,95; RD -0.09; 95% CI -0.14 to -0.04) og deltagelse i sports- eller socialklubber (RR 0,89; 95% CI 0,81–0,98; RD -0.09; 95% CI -0.15 to -0.03). Disse sammenhænge var stærkere blandt kvinder end blandt mænd (Figur 1).

Ift. indlæggelser blev der ikke fundet signifikante forskelle, bortset fra at mænd, der deltog i 3+ aktiviteter (IRR 0,75; 95% CI 0,57–1,00) og mænd der deltog i en religiøs organisation (IRR 0,67; 95% CI 0,46–0,98) havde færre indlæggelser (Figur 2).





Ift. indlæggelser blev der ikke fundet signifikante forskelle, bortset fra at mænd, der deltog i 3+ aktiviteter (IRR 0,75; 95% CI 0,57–1,00), og mænd, der deltog i en religiøs organisation (IRR 0,67; 95% CI 0,46–0,98), havde færre indlæggelser (Figur 2).

Dosis-respons-analyserne viste overordnet et konsistent mønster for personer, der udførte aktiviteterne mindre end én gang om ugen ift. dem der udførte aktiviteterne mindst én gang om ugen. Dog var sammenhængene mellem frivilligt arbejde og lavere dødelighed, samt det at passe en syg eller handicappet og færre indlæggelser, kun til stede for dem, der deltog i aktiviteterne mindre end én gang om ugen. Den modsatte sammenhæng i form af flere indlæggelser var indikeret for personer der passede en syg eller handicappet mere end én gang om ugen <sup>1</sup>.

Sensitivitetsanalyserne viste overordnet robuste resultater. Flere E-værdier viste moderat robusthed. For eksempel blev der fundet en E-værdi på 1,63 for frivilligt arbejde, hvilket betyder, at en umålt confounder, der er associeret 1,63 gange med både eksponeringen og udfaldet, efter at have taget højde for de målte confoundere, vil kunne forklare effekten. Svagere confoundere kan ikke. Den mest robuste E-værdi var for kvinder, der deltog i en religiøs organisation (E-værdi = 3,04, E-værdi nedre CI-grænse = 1,29) <sup>1</sup>.



Når deltagere involveret i religiøse organisationer samt dem med ADL-begrænsninger blev ekskluderet, forblev de overordnede resultater konsistente <sup>1</sup>.

Generelt blev sammenhængene svækket, når alle typer indlæggelser blev inkluderet, og kun sammenhængene i model 1 forblev signifikante <sup>1</sup>.

## **Diskussion**

Denne danske undersøgelse viste, at personer der deltog i mindst én aktivitet, havde lavere dødelighed sammenlignet med personer der ikke deltog i aktiviteter. Specifikt fandt vi, at frivilligt arbejde, det at yde hjælp til familie eller venner samt deltagelse i sports- eller socialklubber var associeret med lavere dødelighed, især blandt kvinder. Mænd, der deltog i mindst tre aktiviteter eller deltog i religiøse organisationer, havde færre indlæggelser. Bemærkelsesværdigt var det, at styrken af sammenhængen for nogle aktiviteter varierede afhængigt af aktivitetsfrekvensen. For mænd var frivilligt arbejde forbundet med en lavere risiko for dødelighed, mens det at passe en syg eller handicappet var forbundet med færre indlæggelser, men kun når aktiviteten blev udført mindre end en gang om ugen. Dette antyder, at overdreven involvering i visse aktiviteter kan ændre sammenhængen fra at være beskyttende til potentielt at have negative sundhedskonsekvenser, i tråd med litteraturen om omsorgsrelateret stress <sup>9</sup>.

Vores resultater stemmer overens med tidligere forskning <sup>8,18-22</sup> som viser, at aktivitetsdeltagelse er associeret med lavere dødelighedsrater. Det interessante er dog, at vores studie indikerer, at deltagelse i sports- eller sociale klubber har en lignende effekt som andre former for socialt engagement, såsom frivilligt arbejde. Dette antyder, at den specifikke type aktivitet muligvis ikke er den primære drivkraft bag de observerede sundhedsfordele. I stedet kan fordelene snarere stamme fra social tilknytning, som kan reducere ensomhed og fungere som en buffer mod stress <sup>23</sup>. Derudover kan deltagelse i sociale aktiviteter give en følelse af mening eller formål med livet <sup>24</sup>, hvilket er forbundet med bedre sundhedseffekter <sup>25</sup>.

Mens kvinder generelt lever længere end mænd på tværs af alle lande <sup>26</sup>, rapporterer de også oftere dårligere selv vurderet helbred, oplever flere depressive symptomer, har en

højere forekomst af samtidige sygdomme, klarer sig dårligere i fysiske tests og er generelt mere skrøbelige end mænd<sup>27-31</sup>. I overensstemmelse med dette fænomen, kendt som "the male-female health survival paradox"<sup>32</sup>, fandt vi, at kvinder drager større fordel af sociale aktiviteter, når det kommer til overlevelse, hvorimod resultaterne indikerer, at selvom den samme sammenhæng ikke findes hos mænd ift. de specifikke aktiviteter, så er det muligt, at aktivitetsdeltagelse kan resultere i færre alvorlige sygdomme og dermed beskytte mænd mod indlæggelser.

Vores tidligere forskning viste, at deltagelse i en religiøs organisation var forbundet med lavere dødelighed for kvinder og færre indlæggelser blandt mænd<sup>11</sup>. Denne undersøgelse fandt lignende sammenhænge for overlevelse, dog lidt mindre udtalte, men ift. indlæggelser, fandt vi ingen lignende sammenhænge for andre aktiviteter. Dette tyder på, at der kan være særlige sundhedsmæssige fordele ved religiøse aktiviteter, især for mænd.

En anden mulig forklaring på vores fund kan være omvendt kausalitet. Deltagelse i sociale aktiviteter kræver både fysisk og mental energi, hvilket betyder, at de, der deltager, muligvis har et bedre helbred end dem, der ikke deltager. Selvom vi justerede for en bred række af potentielle confoundere, herunder medicinsk historie, begrænsninger i dagligdags aktiviteter (ADL) og mental sundhed, som også dækker tidlige tegn på sygdom, såsom mangel på energi og søvnproblemer, kan der stadig være potentielle confoundere, vi ikke har taget højde for f.eks. genetik. Dog reducerer vores longitudinelle studiedesign, hvor aktivitetsdeltagelse blev målt før udfaldet, og hvor der blev taget højde for socioøkonomiske og helbredsmæssige faktorer ved baseline, sandsynligheden for omvendt kausalitet. Derudover viste sensitivitetsanalysen, hvor personer med ADL-begrænsninger ved baseline blev ekskluderet, ingen væsentlige ændringer i resultaterne.

Styrkerne ved dette studie inkluderer et omfattende datasæt, der kombinerer forskellige former for aktivitetsdeltagelse fra interviews med danske registerdata, hvilket giver detaljeret information om dødelighed og indlæggelser over 14 år. Vi undersøgte også dosis-respons-sammenhænge og anvendte både relative og absolutte mål for dødelighed.

Begrænsningerne omfatter en relativt lille studiepopulation på knap 3.000 deltagere, hvilket skaber usikkerhed omkring estimerne. Aktivitetsdeltagelse blev kun målt på ét tidspunkt, hvilket ikke tager højde for ændringer over tid. Forskellige aktiviteter kan have forskellige mekanismer, der påvirker dødelighed, og disse blev ikke fuldt ud undersøgt. En potentiel begrænsning ved SHARE er de lave svarprocenter i nogle lande og interviewbølger, hvilket kan føre til selektionsbias. I Danmark var svarprocenterne 67,1 % i bølge 1 og 65,2 % i bølge 2 <sup>33</sup>.

## Konklusion

Undersøgelsen viste, at deltagelse i mindst én aktivitet var forbundet med lavere dødelighed. Ift. de specifikke aktiviteter var frivilligt arbejde, det at yde hjælp til familie eller venner samt deltagelse i sports- eller socialklubber forbundet med lavere dødelighed, med resultater, der var sammenlignelige med dem for deltagelse i en religiøs organisation. Den stærke sammenhæng med færre indlæggelser blandt mænd, der deltog i en religiøs organisation, blev dog ikke fundet for andre aktiviteter. Fremtidige studier bør undersøge de underliggende mekanismer bag disse sammenhænge.

## Referencer

- 1
- 2 Accepted in Epidemiology, October 14, 2024
- 3 American journal of epidemiology 109  
186 20 1979 [10 1093](#) [11267](#)
- 3 Science (New York, N.Y.) 241 5 0 5 5 1988 [10 1126](#) [3399889](#)
- 5 Appl Psychol Health Well Being 5 28 78 2013  
[10 1111](#) [12000](#)
- 6 PLoS medicine 7 1000316 2010  
[10 1371](#) [1000316](#)
- 7 Psychosomatic medicine 72 229 238 2010  
[10 1097](#) [0 013 3181 01611](#)  
et al.

8 (1982) 251 112906 2020 [10 1016](#) [2020 112906](#) Social science & medicine

9 2020 [10 1002](#) [2 112](#) Campbell Syst Rev 16 112

10 [10 1037 0033 2909 129 6 9 6](#) Psychological bulletin 129 9 6 972 2003

11 Jama 282 2215 2219 1999 [10 1001](#) [282 23 2215](#)

12 epidemiology 38 281 289 2023 [10 1007 1065 023 0096](#) European journal of  
et al.

13 2013 [10 1093](#) [088](#) International journal of epidemiology 42 992 1001

1 Retirement in Europe 202 REGLINK-SHARE - Survey on Health, Ageing and

15 2003 [10 1093](#) [90 1 15](#) Biometrika 90 15 27

[10 1177 1536867 1501500313](#)

16 15 3 809 821 2015

17 [10 7326](#) [16 2607](#) Annals of internal medicine 167 268 27 2017

18 The Stata Journal 20 162 175 2020 [10 1177 1536867 20909696](#)

19 Bmj 367 6377 2019 [10 1136](#) [6377](#)

20 [10 1093](#) [027](#) Health Promot Int 21 5 12 2006

21 589 597 2007 [10 1007 1065 007 9153](#) European journal of epidemiology 22  
et al.

22 community health 70 583 588 2016 [10 1136](#) [2015 206305](#) Journal of epidemiology and  
et al.

23 [10 1016](#) [2009 06 026](#) Preventive medicine 49 1 2 1 7 2009

Journal of health and social behavior 52 1 5 161 2011  
[10 1177 00221 6510395592](#)

2 International Encyclopedia of the Social & Behavioral Sciences  
 (Second Edition) 365 370 2015

25 Health Psychology Review 11 387 18  
 2017 [10.1080/1737199.2017.1327325](#)

26 BMJ 332 808 2006 [332 75 5 808](#)  
 10 1136 332 75 5 808

27 European journal of ageing 17 69 79 2020 [10.1007/10.33.019.0052](#)

28 et al. European journal of public health 29  
 667 67 2019 [10.1093/266](#)

29 Int J Public Health 64 1025 1036  
 2019 [10.1007/00038.019.01270.9](#)

30 Aging & mental health 1 8 2021  
[10.1080/13607863.2021.2013.3](#)

31 Eur. J. Public Health 21 81 91 2011  
[10.1093/022](#)

32 Aging clinical and experimental research  
 20 91 102 2008

33 7 0 0 2019 1 7

## **Triangulation of contradictory evidence from three randomized trials of an early 2-dose measles schedule in Guinea-Bissau from 2003-2019**

**Sebastian Nielsen**<sup>1</sup>, Sören Möller<sup>2</sup>. <sup>1</sup> Bandim Health Project, Department of Clinical Research, University of Southern Denmark, <sup>2</sup> Open Patient data Explorative Network (OPEN), Department of Clinical Research, University of Southern Denmark

*Note: this is work in progress. Suggestions and comments are very welcome! Both before, during and after the symposium, e-mail: [senielsen@health.sdu.dk](mailto:senielsen@health.sdu.dk).*

### **Introduction**

This study investigates the non-specific effects (NSEs) of childhood vaccines on all-cause child mortality in Guinea-Bissau. Nature has recently included the discovery of NSEs of vaccines as their major milestone 13 in the history of vaccine research [1]. In short, vaccines have been shown to alter susceptibility to other disease than the targeted disease, an effect that has been termed NSEs. However, contrary to the specific effects of vaccines, the NSEs are very context dependent. Contextual factors of the NSEs of vaccines include but are not limited to host factors (e.g. sex), sequence and combinations of vaccinations as well as maternal factors [2, 3]. Identification of contextual factors for the NSEs of vaccines is a novel field within the epidemiological investigations of the NSEs of vaccines. Identification of contextual factors often require multiple datasets and triangulation of the available evidence [2, 4].

In this study we propose a method to identify contextual factors, by defining three conditions that needs to be satisfied to identify a contextual factor.

We will use the definition in an explorative investigation with the aim of identifying contextual factors to provide insight on the contradictory results reported from three randomized trials (RCTs) conducted in Guinea-Bissau between 2003 and 2019 [5-7]. RCT I showed a beneficial HR of 0.70 (95% CI: 0.52-0.90) while RCT II produced a HR of 1.38 (0.92-1.90) and RCT III a HR of 1.17 (0.72-1.90) (p-value for homogeneity of HRs = 0.018) (Table 1).

Table 1. Main results and primary characteristics of the three RCTs of 2-dose vs 1-dose measles vaccine (MV) schedules conducted in Guinea-Bissau.

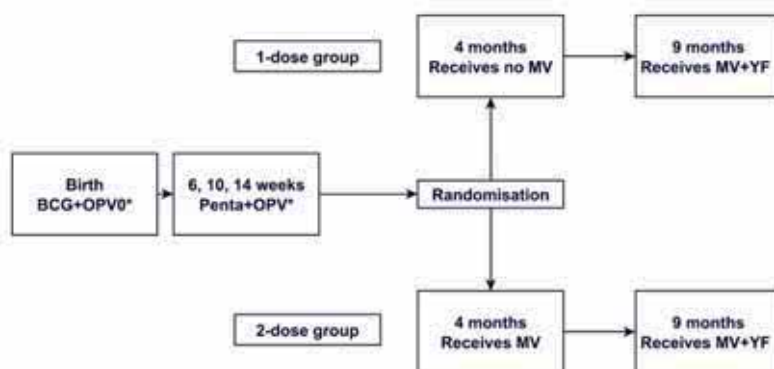
	RCT I	RCT II	RCT III
<b>Hazard ratio (2-dose/1-dose)</b>	0.70 (0.52-0.90)	1.38 (0.92-1.90)	1.17 (0.72-1.90)
<b>Area</b>	Urban	Urban	Rural
<b>Years</b>	2003-2009	2011-2019	2012-2015
<b>Number of participants</b>	6,417	6,598	3,709
<b>Randomization ratio</b>	1:2	2:1	1:1
<b>Number of deaths</b>	217	123	69

## Methods

### Early 2-dose measles vaccine RCTs

The trial design was almost identical for the three RCTs (Figure 1). Children were enrolled from 4 weeks after receiving the third dose of Penta vaccine given from 14 weeks of age. All children received measles vaccine (MV) at 9 months of age as is the World Health Organization (WHO) policy for low-income countries. The RCTs compared the effects on all-cause child mortality of a 2-dose MV schedule given from 4 months and 9 months (2-dose MV group) to the standard one dose of MV at 9 months of age (1-dose MV group).

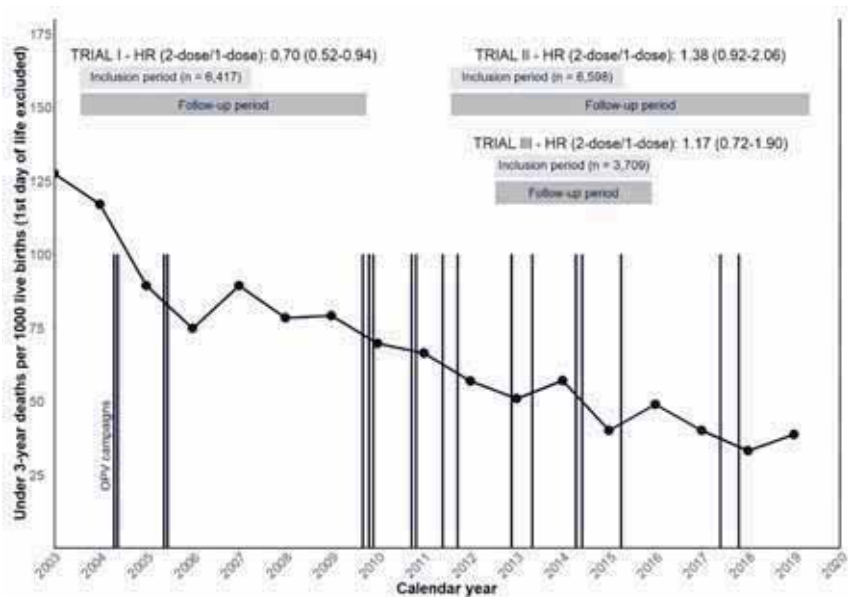
Figure 1. Trial design of the 2-dose MV RCTs.



## Interactions with oral polio vaccination campaigns

The investigators of the RCTs have previously documented a potential interaction affecting the 2-dose/1-dose MV hazard ratio (HR, *unless otherwise stated the HR will always refer to the comparison the 2-dose vs. 1-dose MV group*) with oral polio vaccination campaigns (C-OPV). Children who had received C-OPV-before-enrolment, did not have a beneficial effect of 2-dose MV, compared to the children who did not receive C-OPV-before-enrolment [8, 9]. An overview of enrolment periods, follow-up period and OPV campaigns conducted can be seen in Figure 2.

Figure 2. Inclusion and follow-up period for the three 2-dose MV RCTs in Guinea-Bissau. Vertical lines represent national OPV campaigns conducted.



## Contextual factors

We identified the following potential contextual factors. Non-binary factors were transformed to binary factors as defined below.

*Sequence and combination of vaccines and other health interventions:* C-OPV-before-enrolment (yes / no), C-OPV-after-enrolment (yes / no), age of first routine OPV ( $\leq 14$  days /  $> 14$  days), time interval between DTP3 and enrolment ( $< 5$  weeks /  $\geq 5$  weeks)



and Vitamin A supplementation given at birth (yes / no, OBS: only possible for children in RCT I).

*Host factors:* sex (male / female), age at enrolment (<22 weeks / ≥22 weeks) and exclusive breastfeeding at time of enrolment (yes / no). Maternal antibody (MatAb) measurements (NB: measles MatAb only available for subgroups of RCT I (6.8%) and III (13.4%) and most of RCT II (94.6%). Other MatAb only available in RCT II): detectable measles MatAb (>28.7 mIU/mL / ≤28.7 mIU/mL), rubella (upper quartile / three lower quartiles), varicella (upper quartile / three lower quartiles) and Mumps (upper quartile / three lower quartiles).

### **Triangulation**

In the following we propose a triangulation framework to identify contextual factor(s), which may explain the contradictory results 2-dose MV RCTs. This was done in the combined analysis of all three RCTs for all children and separately by sex since many studies have shown that the NSEs are differential by sex [10].

The triangulation framework consists of three conditions that need to be satisfied:

#### 1) Existence

Contextual factor for the HR exists if either A) the factor independently has a direct significant impact (either positively or negatively) on the HR or B) the factor has a significant impact on the HR combined with sufficiently many two-way interactions with other potential contextual factors (excluding those with factors that are independently impacting the HR as defined in part A). Sufficiently many tests, meaning that the probability of observing the number of significant two-way interactions is less than 5% when adjusted for multiple testing (defined further below) and where the effects of the different two-way-interactions on the HR are all in the same direction.

#### 2) Consistency

The effect of the contextual factor is consistent across RCTs if there is no significant heterogeneity in the impact of the contextual factor on the HR across RCTs.

#### 3) Coherence

The contextual factor is coherent, if the variation in distribution of the contextual factor between RCTs contributes to explaining the differences in HR across RCTs.

If one or more contextual factor(s) can be shown to satisfy these three conditions, by triangulation it is likely that the contextual factor(s) can help explain the observed heterogeneity in the HRs across the RCTs.

### **Beneficial, detrimental or neutral exposure to contextual factors**

It is possible to identify participants who are exposed to multiple contextual factors that affects the HR both beneficially and detrimentally. Participants can spend follow-up time as “only exposed to beneficial contextual factors” (beneficial group), “only exposed to detrimental contextual factors” (detrimental group) or either “exposed to no contextual factors” or “exposed to both beneficial and detrimental contextual factors” (neutral group).

### **Statistical analysis**

We conducted an individual-level combined analysis of the three RCTs, based on the original data sets. For mortality, we present HR and Wald 95% CIs estimated from a Cox proportional hazards model with age as the underlying time and stratified by sex and RCT.

### **Adjusted p-value in multiple testing**

For the existence of a contextual factor by interactions with other factors (two-way interactions, part B in the existence criterion), it is required that the adjusted p-value is less than 5%. We use the cumulative distribution function for the binomial distribution defined as:

$$F(k, n, p) = P(X < k) = \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i}$$

where k is number successes out of n repeated observations of a Bernoulli distributed random variable X with probability p for each success. The formula for the adjusted p-value is then calculated as:

$$(1 - F(k, n, p)) \cdot F(0, n - k, p) \cdot 2$$

The adjusted p-value is the probability to observe k or more successes, i.e. observe at least k significant interactions for a potential contextual factor, but with the requirement that none of the remaining n-k interactions are significant in the opposite direction. Finally, we multiply by two since the direction of the first interaction can be either beneficial or detrimental.

For identifying existence of contextual factors in this study (with  $p=0.025$ ) it translates to a requirement of three ( $k=3$ ) or more significant interactions for both evaluating the contextual factors for all children (12 factors,  $n=22$  tests) and by sex (11 factors,  $n=20$  tests), the corresponding adjusted p-values are 2.1% and 1.7%.

## **Results**

### **Existence**

Among all children the potential beneficial contextual factor “not yet received C-OPV-after-enrolment” was directly impacting the HR. “Not received C-OPV-before-enrolment” was identified as contextual factor with three two-way interactions, namely with: “received OPV at birth by 14 days of age”, “not exclusively breastfeeding” and “female sex”. For the detrimental impact of factors on the HR, no factors were identified to independently significantly impact the HR. With the two-way interactions, three factors were identified as negatively impacting the HR. The first, “received C-OPV-before-enrolment” with four two-way interactions with: “short DTP3 interval”, “detectable measles MatAb”, “received C-OPV-after-enrolment” and “exclusive breastfeeding”. The second, “Received C-OPV-after-enrolment” was impacting negatively in three two-way interactions with: “received C-OPV-before-enrolment”, “male sex” and “high Rubella MatAb levels”. And the third, “Short DTP3 interval” in three two-way interactions with: “received C-OPV-before-enrolment”, “no detectable measles MatAb” and “low varicella MatAb levels”.

Contextual factors identified by sex found among males “not yet received C-OPV-after-enrolment” was an independently beneficial contextual factor. There was no detrimental factor for males. For the females “not received C-OPV-before-enrolment”

was a beneficial factor and “received C-OPV-before-enrolment” was a detrimental contextual factor directly on the HR.

**Consistency**

When comparing the effects of the potential identified contextual factors, we found for all children that the effects among children with beneficial contextual factors were for RCT I-III: 0.76 (0.31-1.83), 0.81 (0.24-2.78) and 1.20 (0.43-3.32) (p for homogeneity = 0.79). For the detrimental factors the HRs were: 2.90 (0.82-10.3), 4.71 (1.43-15.5) and 1.66 (0.68-4.03) (p=0.38). For the contextual factors identified separately by sex both the beneficial and detrimental effects did also not differ significantly across RCTs (p=0.38 and p=0.27).

**Coherence**

In Figure 3 we see that the proportion of follow-up time and the associated mortality rates for the children exposed to beneficial, detrimental or neutral contextual factors as identified for all children by RCT and randomization group. Figure 4 is the same as in Figure 3 but for contextual factors identified separately by sex.

*Figure 3. Proportion and mortality rates by RCT, contextual factor exposure and randomization group for contextual factors identified among all children.*

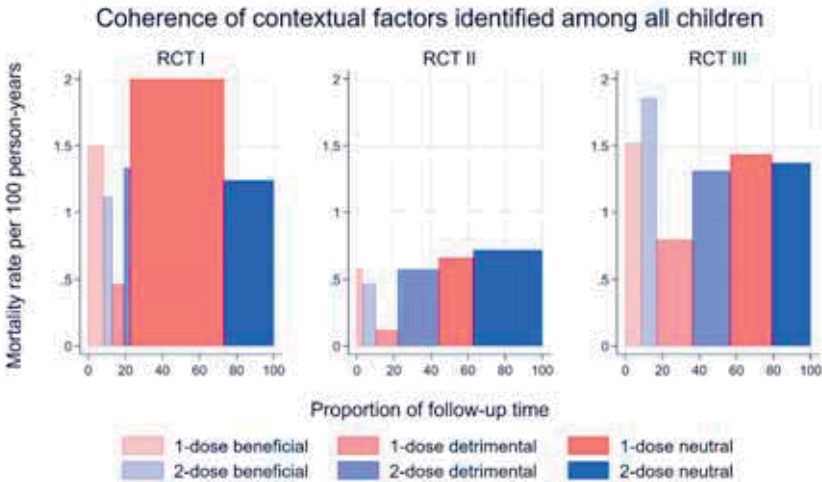
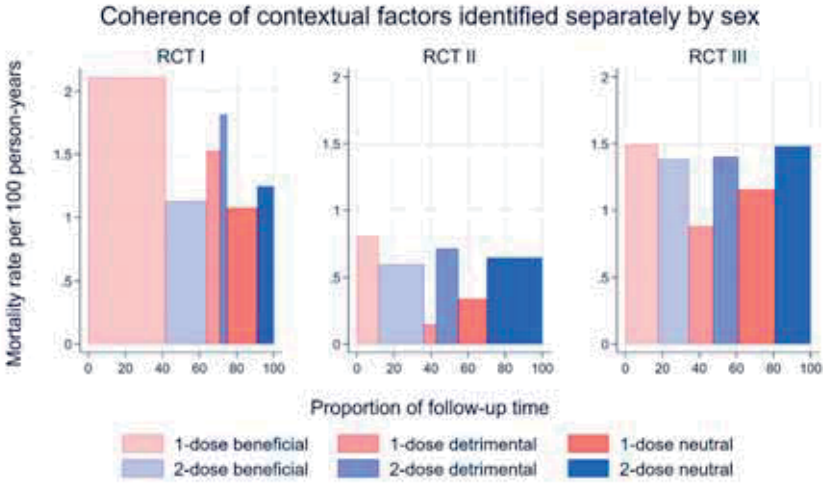


Figure 4. Proportion and mortality rates by RCT, contextual factor exposure and randomization group for contextual factors separately by sex.



Since some of the contextual factors are time-varying the comparison of MRs is only valid between randomization groups (red vs. blue) within contextual exposure groups.

From Figure 3 and 4 we see that the contextual factors identified separately by sex identify a larger proportion of the follow-up time as either beneficial or detrimental, while the contextual factors identified among all children categorizes most follow-up as neutral in all three RCTs.

**Combined HRs by exposure to contextual factors**

The HR based on contextual factors identified among all children exposed to beneficial contextual factors was 0.89 (0.51-1.56) while it was 2.64 (1.47-4.73) among children exposed to detrimental contextual factors (p for same effect = 0.008). The effect was also beneficial in the neutral group, i.e. among children exposed to no contextual factors or exposed to both beneficial and detrimental contextual factors, 0.78 (0.61-0.98). For the contextual factors identified separately by sex the beneficial HR was 0.65 (0.49-0.85) while the detrimental was 1.66 (0.99-2.78) (p for same effect = 0.002) and in the neutral group the effect tended also to be negative, 1.41 (0.97-2.05).

## **Discussion**

Using the proposed definition to identify contextual factors, we corroborated that C-OPV administered both before-enrolment and during follow-up (after-enrolment) impacted the HR (2-dose/1-dose MV) in the three RCTs. For all children two beneficial and three detrimental contextual factors were identified, only some of these were identified separately by sex, mainly those related to C-OPVs. No additional contextual factors were identified by sex. We found that the contextual factors identified separately by sex provided a more coherent explanation of the observed differences in the HR.

## **Strengths and limitations**

We used individual-level data from three RCTs with few differences in trial design and data collection [5-7]. Two RCTs were conducted in the same study area during different time periods (Figure 2). Though RCTs were similar, and data was collected in almost identical manners there may still be differences due to different fieldworkers or procedures in the RCTs. Another limitation is the availability of potential contextual factors in the datasets, e.g. MatAb were only available for some of the participants, while other factors may not be available at all, or we may not be aware that they exist and therefore no information has been collected on these factors. For those factors that are available only some have been chosen to be included in the present triangulation. Notably, we have not (yet) considered whether children “received campaigns with OPV together with vitamin A supplementation” or “campaign Vitamin A supplementation alone”. We did also not distinguish between the comparison of the HR before and after the standard 9-month MV. Before 9-months MV the HR is comparing 1-dose vs no dose of MV and from 9-months MV it is a comparison of 2-dose MV vs. 1-dose MV. There are also several other possible approaches to defining the beneficial and detrimental groups in the RCT cohorts; one approach could be to define a score with beneficial interactions counting positively and detrimental factors negatively. Another aspect would be if all contextual factors are equal, or should they be weighed according to e.g. their level of significance. The cut-off at  $p=0.05$  to identify as potential a contextual factor is arbitrary and could be questioned. Allowing

the inclusion of contextual factors with both beneficial and detrimental effects of the HR, could also be possible with alternative definitions of the framework.

### **Other studies**

We recently investigated the impact of C-OPVs on the HR and found that C-OPVs may have had an impact on the HR and therefore do explain the contradictory results of the three RCTs to some extent [8, 9]. This is however the first study that attempts to define a more rigorous framework for identifying contextual factors for the NSEs of vaccines. Few studies have focused on this very important area of the NSEs.

### **Perspectives**

During the COVID-19 pandemic the Bacillus Calmette–Guérin (BCG) was used in RCTs in both healthcare workers and senior populations as a potential stop-gap vaccine. While BCG vs. placebo reduced all-cause mortality by 39% (3-62%) [11], no RCTs had power to include mortality as a primary outcome, and results on the predefined primary outcomes, e.g. different morbidity outcomes, were mixed. Understanding these effects of BCG and the associated contextual factors would be very important. The framework defined here should be re-tested in other settings and datasets to either confirm, debunk or identify potential new contextual factors associated with not only the comparison of a 2-dose vs. 1-dose MV vaccination schedule, but could also be applied in other areas of epidemiological research.

### **Conclusion**

We defined a triangulation framework and used it to identify the contextual factors to provide insight on the contradictory results of three RCTs of an early 2-dose MV schedule. The findings corroborated previous findings that not having received C-OPV-before-enrolment and C-OPV-after-enrolment affected the HR (2-dose/1-dose) beneficially, while having received C-OPV-before-enrolment, C-OPV-after-enrolment and short DTP3 interval had a detrimental impact on the HR. Despite identified contextual factors varied among all children and separately by sex, and factors identified by sex were more coherent, overall conclusions were similar for both approaches.





## **Estimating Healthcare Transitions: Integrating Logistic Regression and Markov Models to Predict Mortality from Continuity and Discontinuity of Care**

Troels Kristensen, DaCHE, Department of Public Health, University of Southern Denmark

This is a work in progress. Please contact the author for permission before citing or sharing

### **Abstract:**

**Background:** Continuity of care (COC) in general practice has been linked to improved health outcomes. However, maintaining COC becomes increasingly challenging over time due to varying levels of discontinuity. This study conceptualizes COC as a "medical health system technology" with differing degrees of discontinuity (DCOC) across areas, clinics, and subpopulations. The idea is to model the impact of varying effectiveness of this technology on patient mortality.

**Aim:** To model likely transitions from COC to increasing levels of DCOC and related death in patients aged 55 and older in general practice. The intension is to estimate degrees of DCOC over time and enable future applications in resource allocation, care disparities, inequality, and health economics analyses.

**Method:** A Markov model with five states (Continuity, Mild Discontinuity, Moderate Discontinuity, Severe Discontinuity, Death) was used to simulate progression in continuity and cumulative time spent in each discontinuity state towards death. A 12-year cohort of 4,571,041 listed patients above 12 years who stayed alive were applied to measure long run degrees of (D)COC. Annual transition probabilities between states and the absorbing state death were estimated using transition rates, logistic regression analysis and margins analysis at the end of the analysis period. The plan is to use sensitivity analyses assessed the robustness of findings.

**Results:** The model shows transitions across care states based on the estimated annual transition probabilities. This allows for capturing gradual transition and years in continuity of care, mild discontinuity, moderate discontinuity, severe discontinuity and death. The model permits long run cohort simulation for elderly patients. Finally, the model can be used to simulate interventions targeting care coordination reduced transitions to degrees of discontinuity.

**Conclusion:** This type of model offers long run insights into care transitions, enabling evidence-based policy decisions to optimize resource allocation, improve care continuity, and enhance patient outcomes.

## 1. Introduction

Continuity of care (COC) may be conceptualized as a "medical technology" (MedTech) that in most cases positively impacts patient outcomes. While MedTech is typically associated with tangible innovations like drugs, devices, diagnostics, or software - COC or mild levels of discontinuity of care (DCOC) could also be seen as a health system technology. For instance, in general practice. This technology perspective suggests that COC technology shares core characteristics with MedTech. A characteristic is that it relies on structured processes to enable consistent and personalized interactions between patients and providers, delivering measurable benefits such as improved chronic disease management, reduced mortality and cost savings. A well implemented "relation"-technology in terms of COC is widely regarded as a hallmark of an optimal healthcare system, whereas increasing care discontinuity signals systemic dysfunction (i.e., exhibit signs of 'sickness' in health systems). The later deterioration is expected to result in adverse outcomes, including higher mortality rates(1–3).

Using Markov modeling this study intends to actually model the effects of discontinuity in patient care in Danish general practice over time. This means a model where a patient transitions through different states based on their level of COC, subsequent DCOC and outcomes. By doing so, a Markov model can quantitatively assess how COC or degrees of COC impacts progression to outcomes such as death. Potentially also resource use, and quality of life. Given the temporal nature of COC and DCOC, Markov modeling offers a potential logical approach to studying the impact of changes in this "health system technology." Markov models capture transitions between defined states, making them ideal for modeling patient transitions through varying levels of COC, DCOC and their associated health outcomes such as death over time. By simulating transitions to adverse outcomes like mortality, Markov models can estimate cumulative time spent in each state. Next, the impact of interventions to reduce DCOC, and potential gains in patient outcomes could also be analyzed. This study hypothesizes that COC and DCOC can be modelled. The aim of this study is to model likely transitions from COC to increasing levels of discontinuity and related death in patients aged 55 and older in general practice. The intension is to estimate degrees of DCOC over time and enable future applications in resource allocation, care disparities, inequality, and health economics analyses.

## 2. Methods

A Markov model with five states (Continuity, Mild Discontinuity, Moderate Discontinuity, Severe Discontinuity, Death) was used to simulate progression of (dis)continuity (potential disease in health system relations) and cumulative time spent in each state (4). The COC and DCOC states were created based on a combination of statistical and clinical expert opinion. Continuity was defined as zero shifts, mild discontinuity of care was defined as 1-4 shifts. moderate discontinuity was 5-8 shifts, and Severe discontinuity was defined as 9 or more shifts. The empirical measurements of states of progression of discontinuity to death were based on the patients` registered provider shifts over a 12 years analysis period from 2007 to ultimo 2018 (5). The death state was the absorbing state. A cycle length of one 1 year was adopted to allow for gradual transition between states. In principle patients could move between the 5 states in  $5 * 5 = 25$  different ways. The transition matrix (P) specifies all possible probabilities of moving from one state to another. This means the matrix in my Markov model with five states (Continuity(C), Mild Discontinuity(M), Moderate Discontinuity(Mod), Severe Discontinuity(S), Death(D), looks like this:

$$P = \begin{bmatrix} PCC & PCM & PCMod & PCS & PCD \\ PMC & PMM & PMMod & PMS & PMD \\ PModC & PModM & PModMod & PModS & PModD \\ PSC & PSM & PSMoD & PSS & PSD \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

In the transition matrix P in (1),  $p_{ij}$  represents the probability of transitioning from state  $i = C, M, Mod, S, D$  to state  $j = C, M, Mod, S, D$ . The bottom row reflects that death is an absorbing state. The adapted model did not allow for reverse transition probabilities. This means 10 reverse transition probabilities (PMC, PModC, PSC, 0, PModM, PSM, 0, PSMoD, 0, 0) reflecting the zeros below the diagonal in (2) was ruled out. For instance, transition from severe discontinuity back to moderate discontinuity during the analysis period. This was due to the experienced based definition of discontinuity. This means 15 transition probabilities are remaining. They are illustrated in Figure 1 using an illustrative set of transition probabilities.

Figure 1: Illustration of Makrov (dis)continuity states excluding reverse transitions

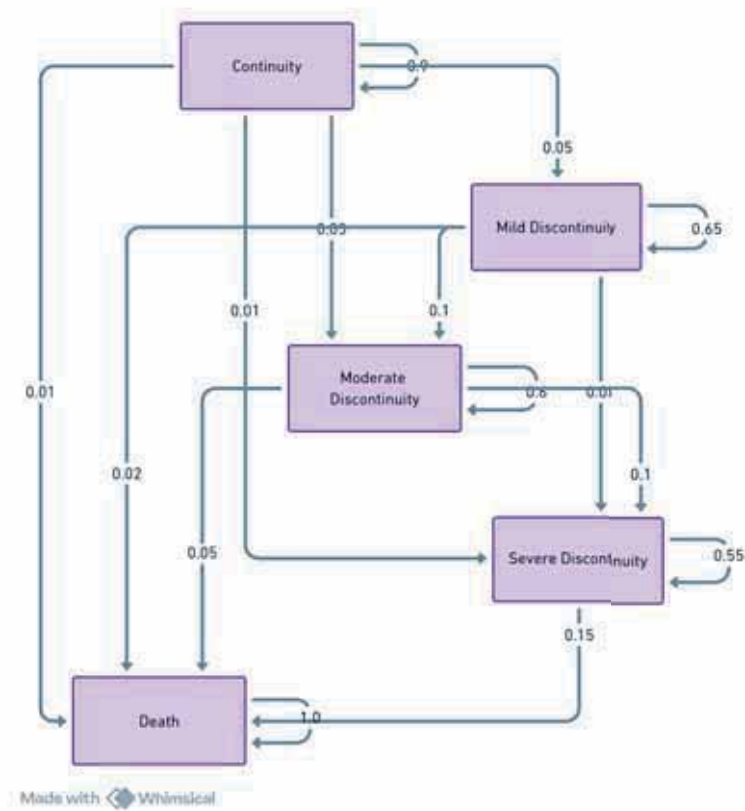


Figure 1 shows an illustrative example of the adapted Markov model without reverse transitions probabilities. They are ruled out since the applied experienced based long run definition of discontinuity does not allow patients to reverse from at state of discontinuity to continuity or lower levels of discontinuity.

Besides the illustrated reduced model, this study anticipates that it is meaningful and reasonable to only model gradual transition (e.g., from continuity to mild, then moderate, and then severe) in one year. This means it is not possible to transition to discontinuity in a way that trespasses the next gradual level. Accordingly, the 3 non-gradual transition

probabilities (PCMod, PCS and PMS) were set to zero. The resultant “reduced transition matrix”  $P^*$  is shown below:

$$P^* = \begin{bmatrix} PCC & PCM & 0 & 0 & PCD \\ 0 & PMM & PMMod & 0 & PMD \\ 0 & 0 & PModMod & PModS & PModD \\ 0 & 0 & 0 & PSS & PSD \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

In total, the transition matrix  $P^*$  in (2) required 11 non-trivial transition probabilities since the zeros and 1 in (2) are given. The 4 diagonal probabilities PCC, PMM, PModMod, PSS of the 11 were calculated as residuals since the probability of staying in the states per definition is the residual up to 1. This means 7 transitions probabilities was required estimated for the model.

The all-cause mortality transition probabilities from the 3 different continuity states (PCD, PMD, PModD) to death, could be assumed to be the same and age dependent based on Danish register data. In contrast, PSD both include the likelihood of Dying from severe continuity plus all-cause mortality.

The implication was that we needed to estimate 3 annual transition probabilities PCM, PMMod, PModS and the later 4 stated “all-cause mortality” probabilities). The resulting Markov states are illustrated in Figure 1. Different from the remaining all-cause mortality probabilities, the transition probability PSD both includes the likelihood of dying from severe discontinuity of care and the likelihood of dying from all-cause mortality.

## 2.1 Transition probabilities

All patients are at risk of gradual transitioning from continuity to mild DCOC, moderate DCOC, severe DCOC at the beginning of analysis period. They also have a risk of all-cause mortality and death from each state. This study used patient-specific register data to estimate transitions probabilities for movements from one state to another within a given time period (cycle) of one year. The applied data originates from a previous and forthcoming

paper(ref). More precisely, the yearly transition probabilities between states in our Markov model was estimated using two main approaches:

***Rate-to-probability:***

The first approach converts transition rates between states into probabilities. The transition rates were calculated based on empirical data.

***Formula:***

If the transition rate from state  $i$  to state  $j$  is  $r_{ij}$ , the probability of transitioning during a cycle of length  $t$  (e.g., one year) is given by:

$$P_{ij} = 1 - e^{-r_{ij}t} \tag{3}$$

To determine Transition Rates  $r_{ij}$  for (3) data was gathered on the number of observed transitions between states ( $i \rightarrow j$ ) and the total time at risk for the population. More precisely, we used data on the Danish population above 12 years listed in Danish general practice to analyze how they switched to the different states during a 12 year analysis period. We Computed the transition rate for the 12 years for (3) as follows:

$$r_{ij} = \frac{\text{Number of Transitions from } i \rightarrow j}{\text{Person-Years at Risk in State } i * 12} \tag{4}$$

Next, we applied the rate-to-probability formula (4) above to convert the transition rate into a probability for 12 years. ( $t = 12$  years).

**2.2 Logistic regression**

These annual transition probabilities to death (PCD, PMD, PModD, PSD) were estimated using logistic regression based on a 12-year cohort of patients aged 50+ and margin analysis.

The dependent variable was a dummy for death in a 3-year follow-up period and the dichotomous explanatory variables was Mild Discontinuity, Moderate Discontinuity, Severe Discontinuity. Logistic regression was used to estimate the probability of an event

(e.g., transitioning from state i to j) as a function of predictors (demographic makers, regional markers, municipality markers, socio economic markers and morbidity markers) at the end of the cohort period ultimo 2018 in the following way:

$$\text{logit}(P_{ij}) = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (5)$$

where  $P_{ij}$  is the transition probabilities and  $X_1, X_2, \dots, X_k$  are the patient characteristic predictors influencing the transition.

***Convert Coefficients to Probabilities:***

For each state the transitioning probabilities were predicted based on the following formula based on the regression in (5):

$$P_{ij} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}$$

More specifically, margins analysis in Stata was used to estimate the required conditional probabilities for the transition matrix. For categorical variables Stata calculates the adjusted probabilities by taking a weighted average of the predicted probabilities across all levels of the categorical variables, using the observed distribution of the variables in the data set. This is usually more realistic than holding variables at their means.

**3. Results**

The results of the logistic regression analysis of the association between follow-up mortality and patient characteristics will be presented at the symposium. This is also the care for the margins analysis of transition probabilities. Below an illustrative example of the model have been included.

**3.1 Transition probabilities for the Markov model**

The illustrative model parameters in Table 1 were used to populate the Markov model:

**Table 1: Gradual transition matrix for Markov model with 5 stages**

	Continuity	Mild discontinuity	Moderate Discontinuity	Severe Discontinuity	Death
Continuity	1-0.05	0.05	0	0	0.01
Mild discontinuity	0	1-0.1	0.1	0	0.02
Moderate Discontinuity	0	0	1-0.1-0.05	0.1	0.05
Severe Discontinuity	0	0	0	1-0.15	0.15
Death	0	0	0	0	1

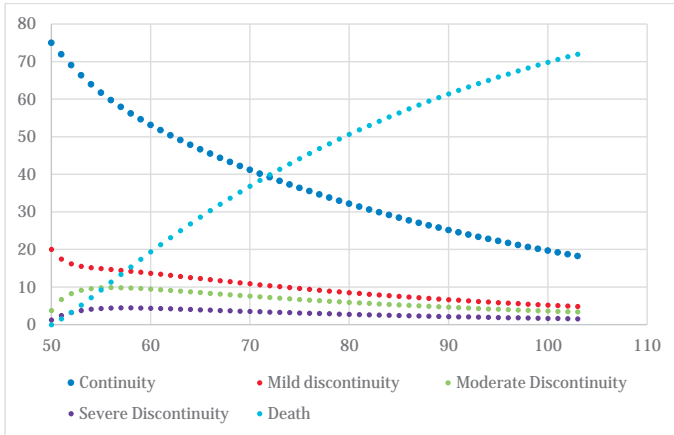
Next in Table 2 the Cohort simulation for the model of COC, degrees of DCOC and death was calculated based on transition probabilities in Table 1 (only the first 10 years for illustration).

**Table 2: Cohort simulation for the model of (dis)continuity transition**

Cycle	Age	Continuity	Mild discontinuity	Moderate Discontinuity	Severe Discontinuity	Death
0	50	75.0	20	3.8	1.3	0.0
1	51	71.9	17.4	6.7	2.4	1.5
2	52	69.0	16.2	8.3	3.2	3.3
3	53	66.3	15.5	9.1	3.8	5.2
4	54	63.9	15.2	9.6	4.1	7.2
5	55	61.7	14.9	9.8	4.3	9.2
6	56	59.7	14.7	9.9	4.4	11.3
7	57	57.9	14.4	9.8	4.5	13.3
8	58	56.2	14.2	9.8	4.4	15.4
9	59	54.6	14.0	9.6	4.4	17.4
10	60	53.1	13.7	9.5	4.4	19.3



**Figure 1: Cohort simulation for elderly patients above 50 years**



#### **4. Discussion:**

This study deals with simulation of gradual transition from COC through likely DCOC to death. It also estimates how much time individuals spend in various states to understand COC and the overall burden of DCOC.

The estimated model or modified versions can be applied to a range of different tasks. One task is healthcare resource planning and workforce allocation. Time in each health state can be used to model the need for healthcare resources, such as improved infrastructure. Another optional task is to adopt the model to assess disparities in care continuity across different populations. In this way, the model can be used simulate the impact of social determinants of health (e.g., income, location) in terms of related transition probabilities on care discontinuity. Furthermore, the model can potentially be used to quantify how interventions targeting underserved populations could reduce transitions to severe discontinuity.

Another potential application is demand forecasting. This means modelling the demand for healthcare services in each state over time, allowing for proactive resource allocation and budgeting. Furthermore, the model can be used assess future Quality of life and conduct health economics analysis. For instance, the model framework can be used to assign health utility scores (e.g., quality-adjusted life years, QALYs) to each state to quantify the impact on patient quality of life over time. It is also an option to attach cost values to each state to calculate the total economic burden of care COC and DCOC over time. Severe discontinuity is expected to have higher associated costs due to likely related hospitalization and lost productivity. Furthermore, attached cost and QALY values to states could allow for cost-effectiveness analysis to determine whether interventions that reduce time in higher-cost states are worth implementing. Our model could also be used to quantify the benefits of policies designed to reduce transitions to higher discontinuity states or assess the impact of specific programs (e.g., care coordination or telemedicine) by modelling and comparing time spent in different states before and after implementation.

### **Limitations:**

#### *Logistic Regression vs. Cox Proportional Hazards Model*

The choice between logistic regression and Cox proportional hazards models depends on the study design and objectives. In this study, logistic regression is used to estimate the probability of mortality during a three-year period following a 12-year cohort period, with predictors based on characteristics measured at the end of the cohort. This design leverages

the cumulative effects of COC and other factors during the cohort period to predict the likelihood of death in the subsequent follow-up period.

Logistic regression is well-suited to this setup because it estimates the probability of death over a fixed follow up period (three years) using predictors that summarize exposures and conditions observed at the end of the cohort period. By focusing on end-of-cohort characteristics, the model simplifies the analysis by avoiding complexities related to time-to-event data or censoring. This approach is particularly effective when the predictors of interest are stable or reflect cumulative effects, such as COC over the preceding 12 years.

However, this approach has some limitations. By using predictors measured at the end of the cohort period, the model does not account for dynamic changes in health status, care patterns, or other critical factors that occurred during the cohort period. These changes may have influenced the predictors or mortality risk in ways that are not explicitly modeled. Furthermore, logistic regression does not capture the timing of deaths within the three-year follow-up period, which could obscure temporal patterns. The Cox proportional hazards model, in contrast, explicitly accounts for time-to-event dynamics, modeling the hazard of death over the follow-up period. It allows for the inclusion of time-varying covariates and captures the timing of deaths, offering a more granular view of mortality risk. However, the Cox model is more complex to implement and may not align as naturally with the discrete transition probabilities required for Markov modeling.

Overall, logistic regression is seen as a pragmatic and effective choice for this study, given the reliance on fixed end-of-cohort characteristics and the clearly defined three-year follow-up period. While it provides probabilities that are readily interpretable and suitable for Markov modeling, complementary analyses using Cox proportional hazards or other survival models could validate these findings and offer additional insights into temporal dynamics and the influence of cumulative and dynamic factors on mortality.

#### *Defining the applied degrees of discontinuity*

In the Danish GP context where there are no specific evidence-based guidelines, defining degrees of discontinuity such as mild, moderate, and severe discontinuity of care. This can be done in several ways. Here we used a combined statistical and conceptual methods. The statistical approach was used to categorize provider shifts based on empirical data distribution. This can ensure objectivity and balanced group sizes. However, statistical thresholds may lack clinical meaning. I also used a more conceptual approach, grounded in

expert consensus among GPs and literature. The later may have aligned the definition with expert knowledge and implications for care quality and patient outcomes such as mortality. But a strict conceptual approach may overlook nuances in population variability or create imbalanced categories. Accordingly, I have attempted to combine these approaches—using both statistical patterns to inform thresholds and refining them with clinical reasoning. Consequently, I hope I have defined discontinuity levels that are both scientifically valid and clinically meaningful.

### References

1. Sandvik H, Hetlevik Ø, Blinkenberg J, Hunskaar S. Continuity in general practice as predictor of mortality, acute hospitalisation, and use of out-of-hours care: a registry-based observational study in Norway. *Br J Gen Pract.* 2022 Feb;72(715):e84–90.
2. Baker R, Freeman GK, Haggerty JL, Bankart MJ, Nockels KH. Primary medical care continuity and patient mortality: a systematic review. *Br J Gen Pract.* 2020 Sep;70(698):e600–11.
3. Pereira Gray DJ, Sidaway-Lee K, White E, Thorne A, Evans PH. Continuity of care with doctors—a matter of life and death? A systematic review of continuity of care and mortality. *BMJ Open.* 2018 Jun 28;8(6):e021161.
4. Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics.* 1998 Apr;13(4):397–409.
5. Olsen J, Kristensen T. Continuity and discontinuity of care among older patients in Danish general practice. *BJGP Open [Internet].* 2023 Jun 17 [cited 2023 Jul 9]; Available from: <https://bjgpopen.org/content/early/2023/06/19/BJGPO.2023.0081>

## Network meta-analysis of diagnostic test accuracy trials

Oke Gerke<sup>1,2</sup>, Werner Vach<sup>3,4</sup>

<sup>1</sup>Department of Nuclear Medicine, Odense University Hospital, Denmark;

<sup>2</sup>Department of Clinical Research, University of Southern Denmark, Denmark; <sup>3</sup>Basel Academy for Quality and Research in Medicine, Switzerland; <sup>4</sup>Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark

### Background

An important step before clinical intervention selection is the diagnosis of the condition of a patient. Diagnostic tests are commonly used to confirm or exclude a target condition (e.g. a disease). To this end, sensitivity (the proportion of positive test results in those with the target condition) and specificity (the proportion of negative test results in those without the target condition) of a diagnostic test are assessed simultaneously. Decision-making rarely relies on a single diagnostic test accuracy study; instead, evidence from multiple diagnostic test accuracy studies addressing the same question is used.

Diagnostic test accuracy meta-analysis focuses on evaluating individual tests across separate studies. However, multiple index tests may be available for a target condition, making their comparative accuracy important for decision-making. Network meta-analysis allows for a more integrated and comprehensive evaluation of several diagnostic tests simultaneously in a single model.

While network meta-analyses are well-established and widely used in outcome studies targeting patient-related benefit like overall survival (1), network meta-analysis in diagnostic accuracy trials differ in terms of simultaneously targeting sensitivity and specificity of a test (2). Several network meta-analysis models have been developed recently to evaluate the comparative accuracy of multiple diagnostic tests, dealing differently with issues like handling of imperfect reference standards, modelling multiple thresholds, or number of diagnostic tests that can be investigated (two vs. more than two) (3).

## Purpose

We performed a systematic review and network meta-analysis on the detection of bone metastases in breast cancer and compared five imaging modalities (4). We applied two meta-analytical approaches: firstly, we applied classical bivariate random-effects modelling (5), thereby making use of all information available from single- and multi-modality studies. Secondly, we focused on the subgroup of paired studies and network meta-analyzed differences in sensitivity and specificity (6). Do single-modality studies provide appropriate added benefit to multi-modality studies?

## Case study

We followed the *Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy studies* (PRISMA-DTA) guidelines (7) and conducted a comprehensive search in MEDLINE/PubMed, Scopus, and Embase. We searched for original research papers in English on humans on 17 March 2023 and updated the search on 25 February 2024. The research question was

*“Does FDG-PET/CT and/or NaF-PET/CT outperform MRI, contrasted enhanced (CE)-CT, and bone scintigraphy in diagnosing bone metastases in breast cancer patients in terms of diagnostic accuracy?”*

We defined PICO(TS) as follows:

- P – Population: high-stage (III or IV) or known primary breast cancer with suspicion of recurrence (staging or re-staging)
- I – Index test(s): imaging with FDG/NaF-PET/CT, MRI, CE-CT, and bone scintigraphy
- C – Comparator: biopsy or clinical / imaging follow-up (composite reference)
- O – Outcome: summary sensitivity & summary specificity (patient-based)
- T – Timing: no limit on date of publication (but inherently limited for PET/CT)
- S – Study type: original clinical studies

All studies included in the meta-analysis needed to have information on both sensitivity and specificity on a per-patient basis. Studies with missing information on

either target parameter were excluded and missing information was not imputed. We conducted all analyses with the package *metadta* and the *network* suite in STATA/BE 18.0 (StataCorp, College Station, Texas 77845 USA).

## Results

### *Study selection*

We identified 913 records through database search and assessed 48 full-text articles for eligibility. Forty studies contributed with data to the systematic review (Figure 1), of which 29 studies contributed to our meta-analyses. Reasons for exclusion from meta-analysis were missing information on specificity, missing information on prevalence of bone metastases, lesion-based analysis only, and half of the patients (5 out of 9) receiving chemotherapy prior to imaging.



Figure 1. Regional distribution of 40 studies (on 4,247 patients) published between 2001 and 2023

### *Study characteristics*

Half of the studies were retrospective (20 out of 40, 50%), 17 were prospective (42.5%), and three were unclear (7.5%). Number of patients (bone metastases) ranged from 9–884 (0–95), with a median and interquartile range of 85 and 51–119

(27 and 12–42), respectively. Of the 29 studies that contributed to our quantitative synthesis of results investigated 13 studies (45%) a single modality (7 times FDG-PET/CT, 5 times NaF-PET/CT, and once CE-CT) and 16 studies (55%) assessed at least two modalities simultaneously (Figure 2, from left to right):

- once FDG-PET/CT, MRI, CE-CT, and bone scintigraphy.
- once FDG-PET/CT, NaF-PET/CT, and bone scintigraphy,
- once FDG-PET/CT, CE-CT, and bone scintigraphy,
- once FDG-PET/CT and NaF-PET/CT,
- once FDG-PET/CT and CE-CT,
- 8 times FDG-PET/CT and bone scintigraphy,
- once MRI and CE-CT, and, finally,
- twice MRI and bone scintigraphy.

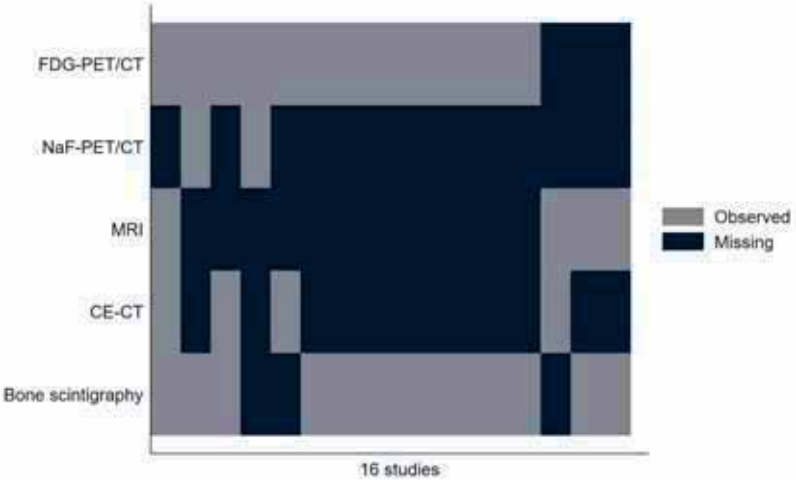


Figure 2. Pattern of which imaging modalities were compared in the included studies [reproduced from (4) with permission].



Figure 3 shows the respective network map, emphasizing that FDG-PET/CT and bone scintigraphy were the most frequently compared modalities, followed by FDG-PET/CT and CE-CT as well as MRI and bone scintigraphy.

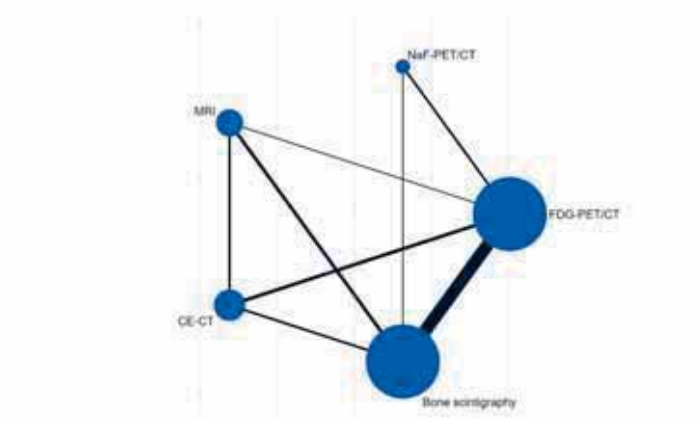


Figure 3. Network map [reproduced from (4) with permission]

*Bivariate random-effects modelling*

Figure 4 comprises Summary Receiver-Operating-Characteristics (SROC) curves derived from bivariate random effects models by modality. The generalized I-square statistic for between-study heterogeneity assessment was 35.64%, 0.03%, 0.01%, 0%, and 45.13% for FDG-PET/CT, NaF-PET/CT, MRI, CE-CT, and bone scintigraphy, respectively. So, notable between-study heterogeneity was observed for both FDG-PET/CT and bone scintigraphy.

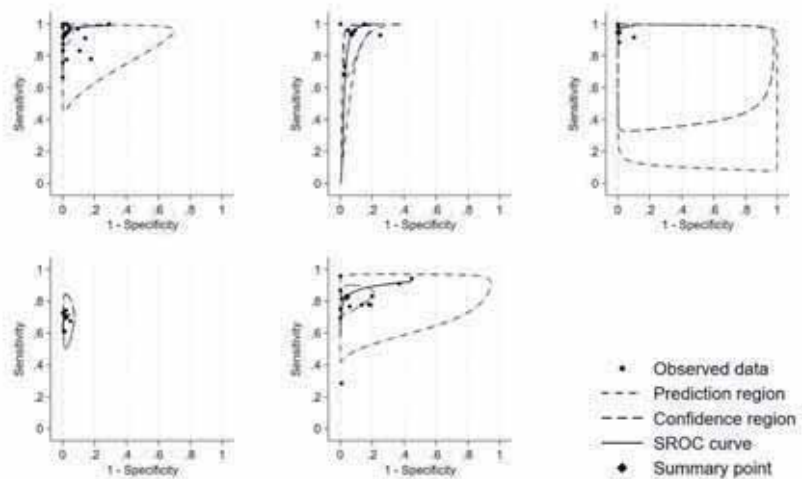
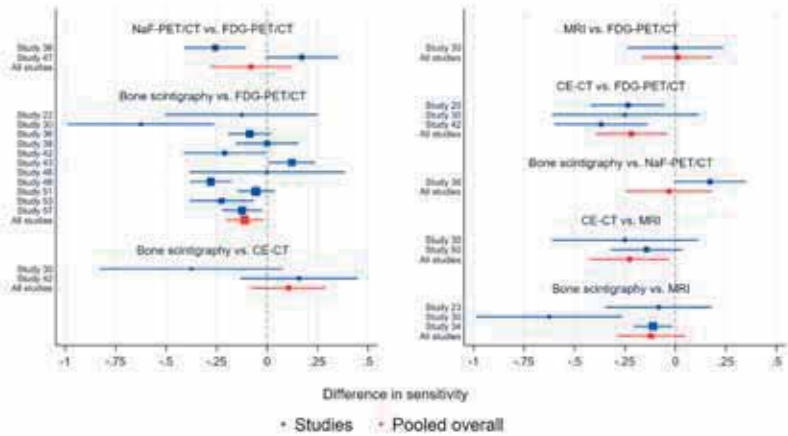


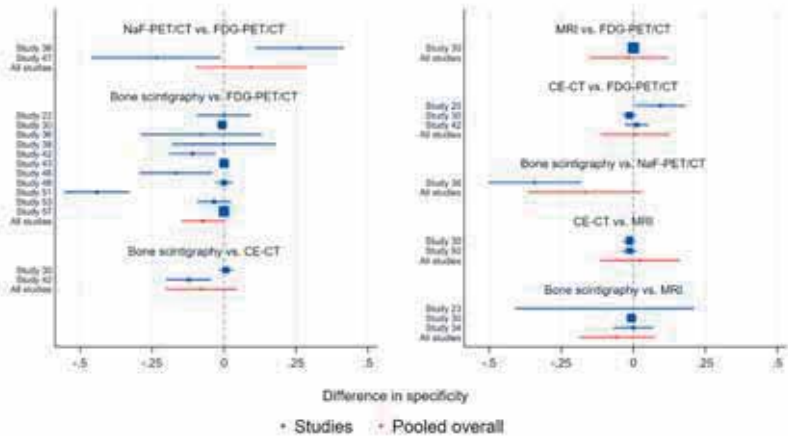
Figure 4. SROC curves for FDG-PET/CT (top left), NaF-PET/CT (top center), MRI (top right), CE-CT (bottom left), bone scintigraphy (bottom right) [reproduced from (4)]

### *Network meta-analysis*

Figure 5 visualizes the estimated differences in sensitivity and specificity for all estimable comparisons between modalities (as shown in Figure 3). It is apparent that the two studies that compared FDG-PET/CT and NaF-PET/CT directly were diametrical in terms of superiority in sensitivity or specificity, suggesting differing thresholds for PET positivity. Comparing FDG-PET/CT and bone scintigraphy, estimates for the difference in sensitivity varied a lot. Those for the difference in specificity were consistently zero for 6 out of 11 studies, but advantageous for FDG-PET/CT for the remaining studies. The difference in specificity between MRI and CE-CT was quite precisely estimated in two studies whereas the summary estimate became comparably imprecise (in terms of a comparably wide 95% confidence interval).



Graphs by column



Graphs by column

Figure 5. Estimated differences in sensitivity (top) and specificity (bottom) for all estimable pairwise comparisons of modalities. Positive values indicate an advantage for the modality named first in the comparison. [Reproduced from (4) with permission]

Differences between estimates from direct comparisons and indirect comparisons exceeded 20% for differences in both sensitivity and specificity for several head-to-head comparisons (e.g. FDG-PET/CT vs. NaF-PET/CT; NaF-PET/CT vs. bone scintigraphy), see Table 1. This suggests an inconsistency issue for some pairwise comparisons of modalities.

Table 1: Sidesplits in network meta-analyses for differences in sensitivity (A) and specificity (B) [reproduced from (4) with permission]

(A) Sensitivity

Side	Direct comparison		Indirect comparison		Difference		
	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	P-value
<b>FDG-PET/CT vs. NaF-PET/CT</b>	-0.05	0.11	-0.29	0.29	<b>0.24</b>	0.30	0.42
FDG-PET/CT vs. MRI	0	0.18	0.01	0.11	-0.01	0.21	0.95
<b>FDG-PET/CT vs. CE-CT</b>	-0.29	0.10	0	0.17	<b>-0.29</b>	0.19	0.13
<b>FDG-PET/CT vs. BS</b>	-0.12	0.05	0.11	0.20	<b>-0.23</b>	0.21	0.27
<b>NaF-PET/CT vs. BS</b>	0.16	0.15	-0.19	0.14	<b>0.35</b>	0.20	0.086
MRI vs. CE-CT	-0.20	0.13	-0.27	0.17	0.07	0.21	0.74
<b>MRI vs. BS</b>	-0.20	0.10	0.10	0.18	<b>-0.30</b>	0.21	0.16
<b>CE-CT vs BS</b>	-0.04	0.16	0.18	0.12	<b>-0.22</b>	0.20	0.27

(B) Specificity

Side	Direct comparison		Indirect comparison		Difference		
	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	P-value
<b>FDG-PET/CT vs. NaF-PET/CT</b>	0.05	0.11	0.34	0.28	<b>-0.29</b>	0.31	0.35
FDG-PET/CT vs. MRI	0	0.13	-0.03	0.09	0.03	0.15	0.87
FDG-PET/CT vs. CE-CT	0.03	0.07	-0.07	0.13	0.09	0.15	0.53
FDG-PET/CT vs. BS	-0.07	0.04	-0.13	0.18	0.06	0.18	0.72
<b>NaF-PET/CT vs. BS</b>	-0.37	0.13	0.05	0.14	<b>-0.42</b>	0.20	0.037
MRI vs. CE-CT	-0.01	0.09	0.11	0.14	-0.13	0.16	0.43
MRI vs. BS	-0.02	0.08	-0.16	0.13	0.14	0.15	0.36
<b>CE-CT vs. BS</b>	-0.05	0.09	-0.11	0.10	0.06	0.13	0.67

Table 2 shows estimates and 95% confidence intervals of sensitivity and specificity for each modality, derived from both single- and multi-modality studies. With a sensitivity of around 0.94-0.95 and a specificity between 0.93 and 1, FDG-PET/CT, MRI, and NaF-PET/CT outperformed CE-CT and bone scintigraphy. The network meta-analysis of multi-modality studies supported the comparable performance of FDG-PET/CT and MRI in the assessment of bone metastases whereas the contrary estimated differences in sensitivity and specificity between FDG-PET/CT and NaF-PET/CT (-8% vs. +10%) are inconsistent with the respective point estimates of the two modalities. For FDG-PET/CT, MRI, CE-CT, and bone scintigraphy, the estimates of sensitivity and specificity from bivariate random effects modelling and the estimated differences in sensitivity and specificity between modalities from the network meta-analysis are consistent.

Table 2 Summary of meta-analyses. Estimates (and respective 95% confidence intervals) of sensitivity and specificity derived from bivariate random effects models and network meta-analyzed differences in sensitivity and specificity

Modality	Bivariate random effects models			Network meta-analysis of multi-modality studies (N=16); comparator minus reference	
	Number of studies	Sensitivity	Specificity	Sensitivity	Specificity
FDG-PET/CT	20	0.94 (0.89–0.97)	0.98 (0.96–0.99)	Reference	Reference
MRI	7	0.94 (0.82–0.98)	0.93 (0.87–0.96)	0.01 (-0.16 – 0.18) p=0.91	-0.02 (-0.15 – 0.12) p=0.82
NaF-PET/CT	4	0.95 (0.85–0.98)	1 (0.93–1)	-0.08 (-0.28 – 0.12) p=0.44	0.10 (-0.10 – 0.29) p=0.34
CE-CT	5	0.70 (0.62–0.77)	0.98 (0.97–0.99)	-0.22 (-0.40 – -0.04) p=0.017	0.01 (-0.11 – 0.13) p=0.92
Bone scintigraphy	13	0.83 (0.75–0.88)	0.96 (0.87–0.99)	-0.11 (-0.20 – -0.02) p=0.017	-0.07 (-0.15 – 0.001) p=0.053

## Conclusions

Our twofold analysis strategy indicated clearly consistent as well as inconsistent pairwise comparisons between the five modalities. Both FDG-PET/CT and MRI accurately assessed bone metastases in breast cancer in terms of sensitivity and specificity, and they outperformed CE-CT and bone scintigraphy regarding sensitivity. On the opposite, the potential role of NaF-PET/CT here is still unclear. Earlier critiques advised against performing single diagnostic test evaluation at all (8). Using single-modality studies in network meta-analyses of diagnostic test accuracy trials should be done with caution (if at all), and further research and exploration into caveats are warranted.

## References

- (1) Cochrane Handbook for Systematic Reviews of Interventions. Chapter 11: network meta-analyses. <https://training.cochrane.org/handbook>
- (2) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. <https://training.cochrane.org/handbook-diagnostic-test-accuracy>
- (3) Veroniki AA, Tsokani S, Agarwal R, Pagkalidou E, Rucker G, Mavridis D, Takwoingi Y. Diagnostic test accuracy network meta-analysis methods: A scoping review and empirical assessment. *J Clin Epidemiol.* 2022; 146: 86-96. <https://doi.org/10.1016/j.jclinepi.2022.02.001>
- (4) Gerke O, Naghavi-Behzad M, Nygaard ST, Sigaroudi VR, Vogsen M, Vach W, Hildebrandt MG. Diagnosing Bone Metastases in Breast Cancer: A Systematic Review and Network Meta-Analysis on Diagnostic Test Accuracy Studies of 2-[18F]FDG-PET/CT, 18F-NaF-PET/CT, MRI, Contrast-Enhanced CT, and Bone Scintigraphy. *Semin Nucl Med.* 2025 Jan;55(1):137-151. <https://doi.org/10.1053/j.semnuclmed.2024.10.008>
- (5) Nyaga VN, Arbyn M. Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data - a tutorial. *Arch Public Health.* 2022; 80(1): 95. <https://doi.org/10.1186/s13690-021-00747-5>
- (6) White IR. Network meta-analysis. *Stata Journal.* 2015; 15(4): 951-985. <https://doi.org/10.1177/1536867X1501500403>
- (7) McInnes MDF, Moher D, Thoms BD, McGrath TA, Bossuyt PM; and the PRISMA-DTA Group; Clifford T, Cohen JF, Deeks JJ, Gatsonis C, Hooft L, Hunt HA, Hyde CJ, Korevaar DA, Leeflang MMG, Macaskill P, Reitsma JB, Rodin R, Rutjes AWS, Salameh JP, Stevens A, Takwoingi Y, Tonelli M, Weeks L, Whiting P, Willis BH. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA.* 2018 Jan 23;319(4):388-396. <https://doi.org/10.1001/jama.2017.19163>
- (8) Moons KG, van Es GA, Michel BC, Büller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. *Epidemiology.* 1999; 10(3): 276-281.