Machine learning (ML) til sikring af datakvalitet i de nationale sundhedsregistre

Sundhedsdatastyrelsen (SDS) d. 9. oktober 2024

E-observatoriet 2024

Mikkel Wermer Steen

SUNDHEDSDATA-STYRELSEN

Indhold

- **Baggrund, resultater og erfaringer med ML-initiativer til forbedring af datakvalitet**
- **Planer og fremtidigt perspektiv**
- **Afrunding / spørgsmål**

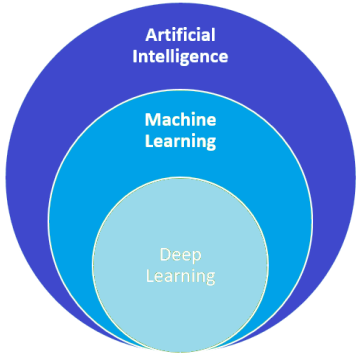
Baggrund og formål med ML-datakvalitetsprojekter

Hvad ser vi i SDS?

1. **Mere kompleks data:**
   - Vi modtager løbende et større og mere varieret volumen af data
   - Datastrukturen bliver stadig mere kompleks

2. **Øget interesse for sundhedsdata:**
   - Flere ønsker adgang til data
   - Gælder både til primær og sekundær anvendelse

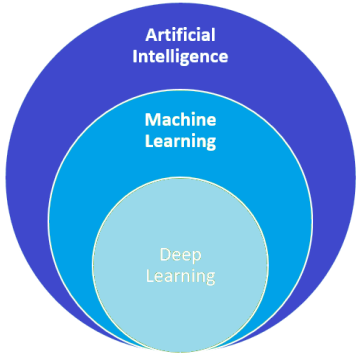
3. **Nye initiativer på vej:**
   - EHDS-forordningen er under implementering
   - Nationalt set ønskes data mere tilgængeligt
   - Fokus er på bedre deling og tilgængelighed af data



Hvordan påvirker det SDS?

1. **Større krav:**
   - Vi oplever stigende efterspørgsel på oplysninger om data:
     - Tilgængelighed
     - Indhold
     - Datakvalitet

2. **Behov for ibrugtagen af nye metoder til:**
   - Videreudvikling og effektivisering af det eksisterende datakvalitetsarbejde
   - Kvalitetsarbejde med nuværende og fremtidige centrale sundhedsregistre



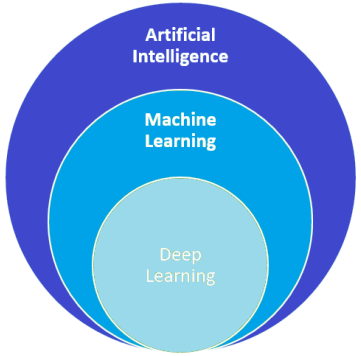
Hvad gør SDS?

AI-metoder afprøves med henblik på at:

1. **Identificere fejl og mangler i data**
   - Se skjulte mønstre i vores sundhedsregistre
   - Kunne fange og reagere hvis noget ændrer sig i måden, data indberettes på

2. **Få overblik over vores data**
   - Finde ud af i hvilke situationer der typisk forekommer fejl eller mangler

3. **Opdage problemer tidligt**
   - Finde fejl, før de bliver et stort problem
   - Hurtigt kunne gribe ind med kvalitetsforbedrende tiltag



Anvendelse af ML til at understøtte datakvalitetsarbejdet

Succeskriterier for datakvalitetsinitiativerne

Forbedre datakvalitet

- Identificere ændringer i indberetningsmønstre og indhold
- Afdække fejl og mangler i sundhedsdata
- Få overblik over datakvalitet på tværs af datakilder

AI kompetenceudvikling

- Opbygge kompetencer internt
- Sidemandsoplæring og gennemsigtighed (undgå ”black box”)
- Udvikling af generiske metoder som kan benyttes på tværs af flere indberetningsområder

Datakvalitetsarbejde i SDS

SDS har undersøgt, hvordan ML kan forbedre kvaliteten af sundhedsdata.

Case 1
Cancer diagnoser

Validering af diagnoseindberetninger til Landspatientregistret

Pilotprojekt afsluttet, plan for videreførelse af projektet

Case 2
Outlier detection

Identificering af anomalier i indberetningsfrekvenser i Landspatientregistret

Pilotprojekt afsluttet, plan for videreførelse af projektet

Case 3
Dødsårsager

Validering af dødsårsagskoder i Dødsårsagsregistret

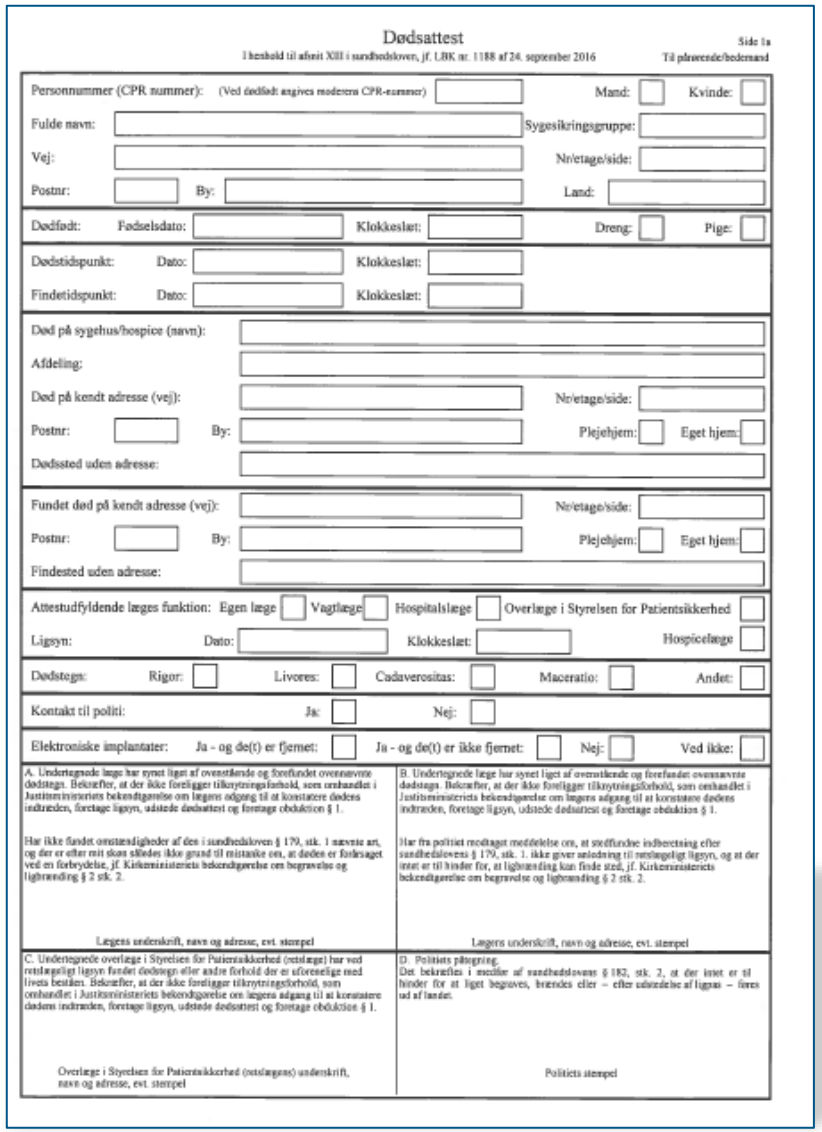
Modeludvikling i afsluttende fase, indledende test ift. idriftsættelse er i gang

Fokus i dag er på validering af dødsårsagskoder (Case 3) i Dødsårsagsregisteret (DAR)

Case 3 – Dødsårsager

Proces for bestemmelse af den underliggende dødsårsag

Borger dør



Dødsattest

I henhold til afsnit XIII i sundhedsloven, jf. LBK nr. 1188 af 24. september 2016

Side 1a

Til pårørende/bedemand

Personnummer (CPR nummer): (Ved dødfødt: angives moderens CPR-nummer) Mand: Kvinde:

Fulde navn: Sygesikringsgruppe:

Vej: Nr/etage/side:

Postnr: By: Land:

Dødfødt: Fødselsdato: Klokkeslæt: Dreng: Pige:

Dødstidspunkt: Dato: Klokkeslæt:

Findetidspunkt: Dato: Klokkeslæt:

Død på sygehus/hospice (navn):

Afdeling:

Død på kendt adresse (vej): Nr/etage/side:

Postnr: By: Plejehjem: Eget hjem:

Dødssted uden adresse:

Fundet død på kendt adresse (vej): Nr/etage/side:

Postnr: By: Plejehjem: Eget hjem:

Findested uden adresse:

Attestudfyldende læges funktion: Egen læge Vagtlæge Hospitalslæge Overlæge i Styrelsen for Patientsikkerhed

Ligsyn: Dato: Klokkeslæt: Hospicelæge

Dødstegn: Rigor: Livores: Cadaverositas: Maceratio: Andet:

Kontakt til politi: Ja: Nej:

Elektroniske implantater: Ja - og de(t) er fjernet: Ja - og de(t) er ikke fjernet: Nej: Ved ikke:

A. Undertegnede læge har synet liget af ovenstående og forefundet ovennævnte dødstegn. Bekræfter, at der ikke foreligger tilknytningsforhold, som omhandlet i Justitsministeriets bekendtgørelse om lægens adgang til at konstatere dødens indtræden, foretage ligsyn, udstede dødsattest og foretage obduktion § 1.

Har ikke fundet omstændigheder af den i sundhedsloven § 179, stk. 1 nævnte art, og der er efter mit skøn således ikke grund til mistanke om, at døden er forårsaget ved en forbrydelse, jf. Kirkeministeriets bekendtgørelse om begravelse og ligbrænding § 2 stk. 2.

Lægens underskrift, navn og adresse, evt. stempel

B. Undertegnede læge har synet liget af ovenstående og forefundet ovennævnte dødstegn. Bekræfter, at der ikke foreligger tilknytningsforhold, som omhandlet i Justitsministeriets bekendtgørelse om lægens adgang til at konstatere dødens indtræden, foretage ligsyn, udstede dødsattest og foretage obduktion § 1.

Har fra politiet modtaget meddelelse om, at stedfundne indberetning efter sundhedslovens § 179, stk. 1. ikke giver anledning til retslægeligt ligsyn, og at der intet er til hinder for, at ligbrænding kan finde sted, jf. Kirkeministeriets bekendtgørelse om begravelse og ligbrænding § 2 stk. 2.

Lægens underskrift, navn og adresse, evt. stempel

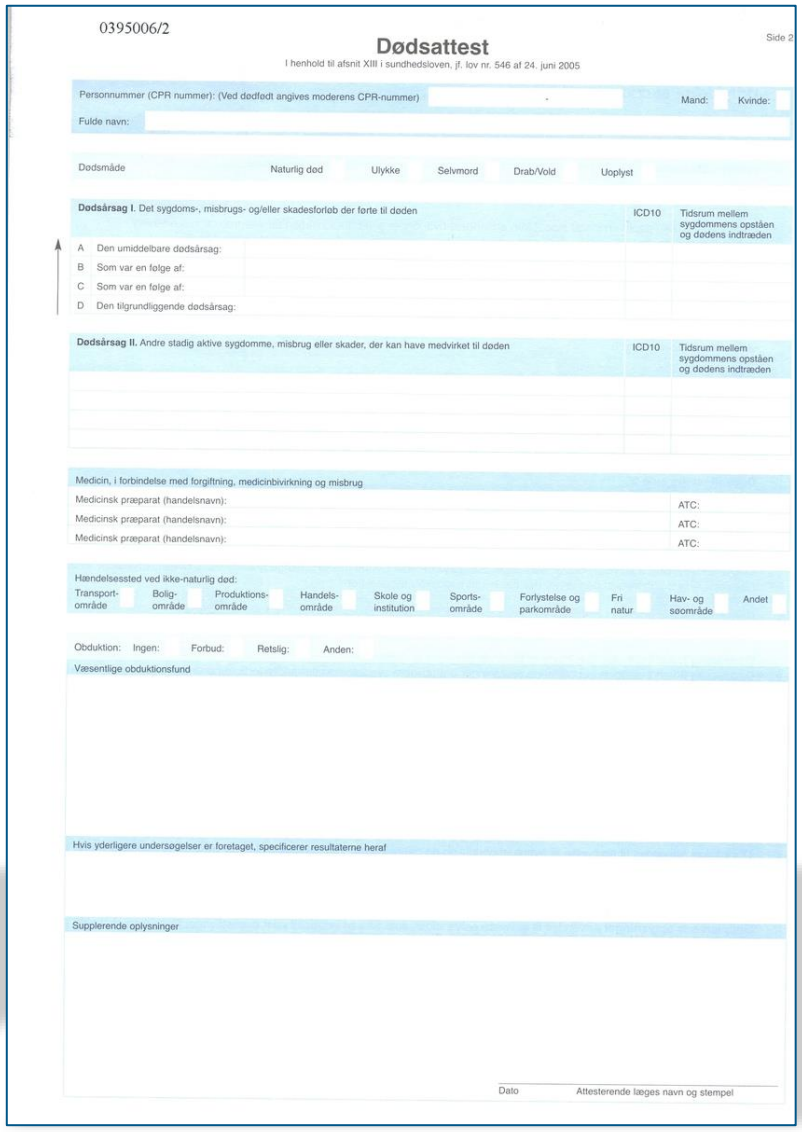
C. Undertegnede overlæge i Styrelsen for Patientsikkerhed (retslæge) har ved retslægeligt ligsyn forårsat dødstegn eller andre forhold der er uforenelige med livets bestået. Bekræfter, at der ikke foreligger tilknytningsforhold, som omhandlet i Justitsministeriets bekendtgørelse om lægens adgang til at konstatere dødens indtræden, foretage ligsyn, udstede dødsattest og foretage obduktion § 1.

Overlæge i Styrelsen for Patientsikkerhed (retslægens) underskrift, navn og adresse, evt. stempel

D. Politiets påtegning.
Det bekræftes i medfør af sundhedslovens § 182, stk. 2, at der intet er til hinder for at liget begraves, brændes eller – efter udstedelse af ligpas – føres ud af landet.

Politiets stempel

Læge udfylder dødsattestens side 1



0395006/2

Dødsattest

I henhold til afsnit XIII i sundhedsloven, jf. lov nr. 546 af 24. juni 2005

Side 2

Personnummer (CPR nummer): (Ved dødfødt angives moderens CPR-nummer) - Mand: Kvinde:

Fulde navn:

Dødsmåde Naturlig død Ulykke Selvmord Drab/Vold Uoplyst

Dødsårsag I. Det sygdoms-, misbrugs- og/eller skadesforløb der førte til døden ICD10 Tidsrum mellem sygdommens opståen og dødens indtræden

A Den umiddelbare dødsårsag:

B Som var en følge af:

C Som var en følge af:

D Den tilgrundliggende dødsårsag:

Dødsårsag II. Andre stadig aktive sygdomme, misbrug eller skader, der kan have medvirket til døden ICD10 Tidsrum mellem sygdommens opståen og dødens indtræden

Medicin, i forbindelse med forgiftning, medicinbivirkning og misbrug

Medicinsk præparat (handelsnavn): ATC:

Medicinsk præparat (handelsnavn): ATC:

Medicinsk præparat (handelsnavn): ATC:

Hændelsessted ved ikke-naturlig død:

Transport-område Bolig-område Produktions-område Handels-område Skole og institution Sports-område Forlystelse og parkområde Fri natur Hav- og søområde Andet

Obduktion: Ingen: Forbud: Retslig: Anden:

Væsentlige obduktionsfund

Hvis yderligere undersøgelser er foretaget, specificerer resultaterne heraf

Supplerende oplysninger

Dato Attesterende læges navn og stempel

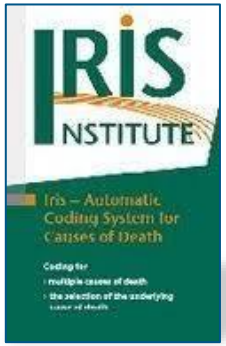
Læge udfylder dødsattestens side 2

Proces for bestemmelse af den underliggende dødsårsag

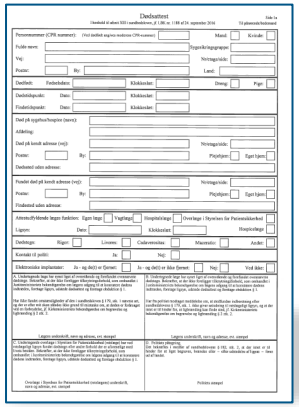
Borger dør

Indberetning til SDS

Bestemmelse af dødsårsag (IRIS)



DAR



Læge udfylder dødsattestens side 1 og 2

Indlæsning/databehandling

Fritekst-information

Ulykker

Bredere datagrundlag

Erfaringsgrundlag

CODA
Cause Of Death Analysis

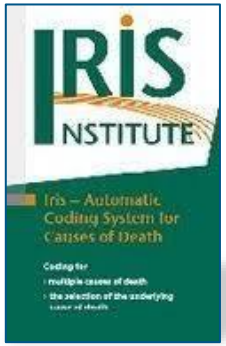
Manuel kodning af dødsårsager hvor der bl.a. anvendes eksterne datakilder som Landspatientregisteret + Patologiregisteret

Proces for bestemmelse af den underliggende dødsårsag

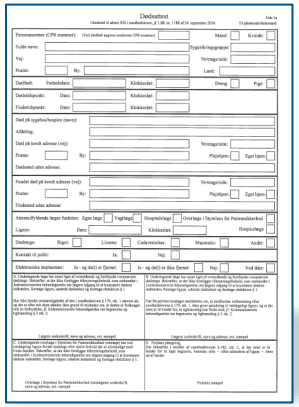
Borger dør

Indberetning til SDS

Bestemmelse af dødsårsag (IRIS)

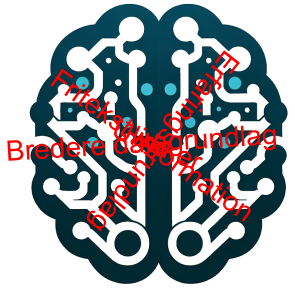


DAR

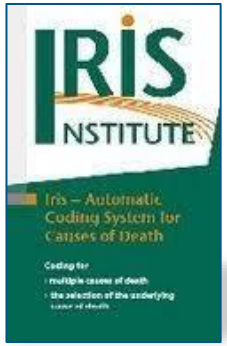


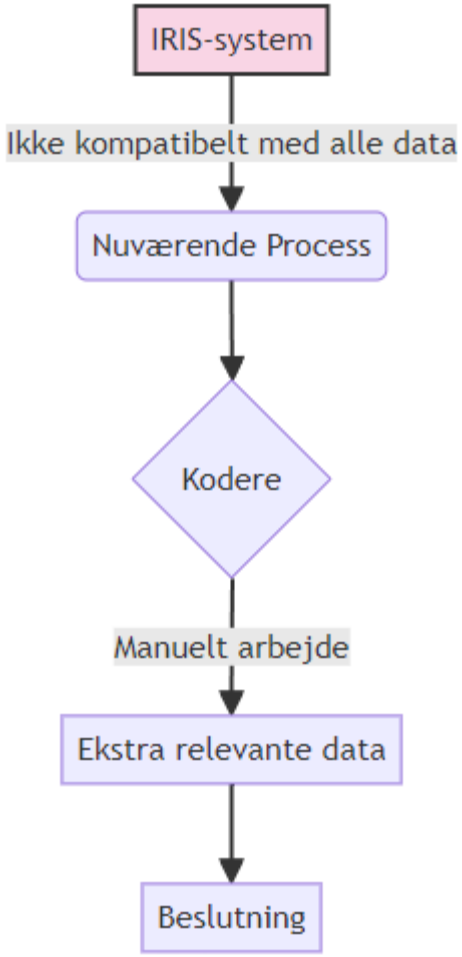
Læge udfylder dødsattestens side 1 og 2

Indlæsning/databehandling

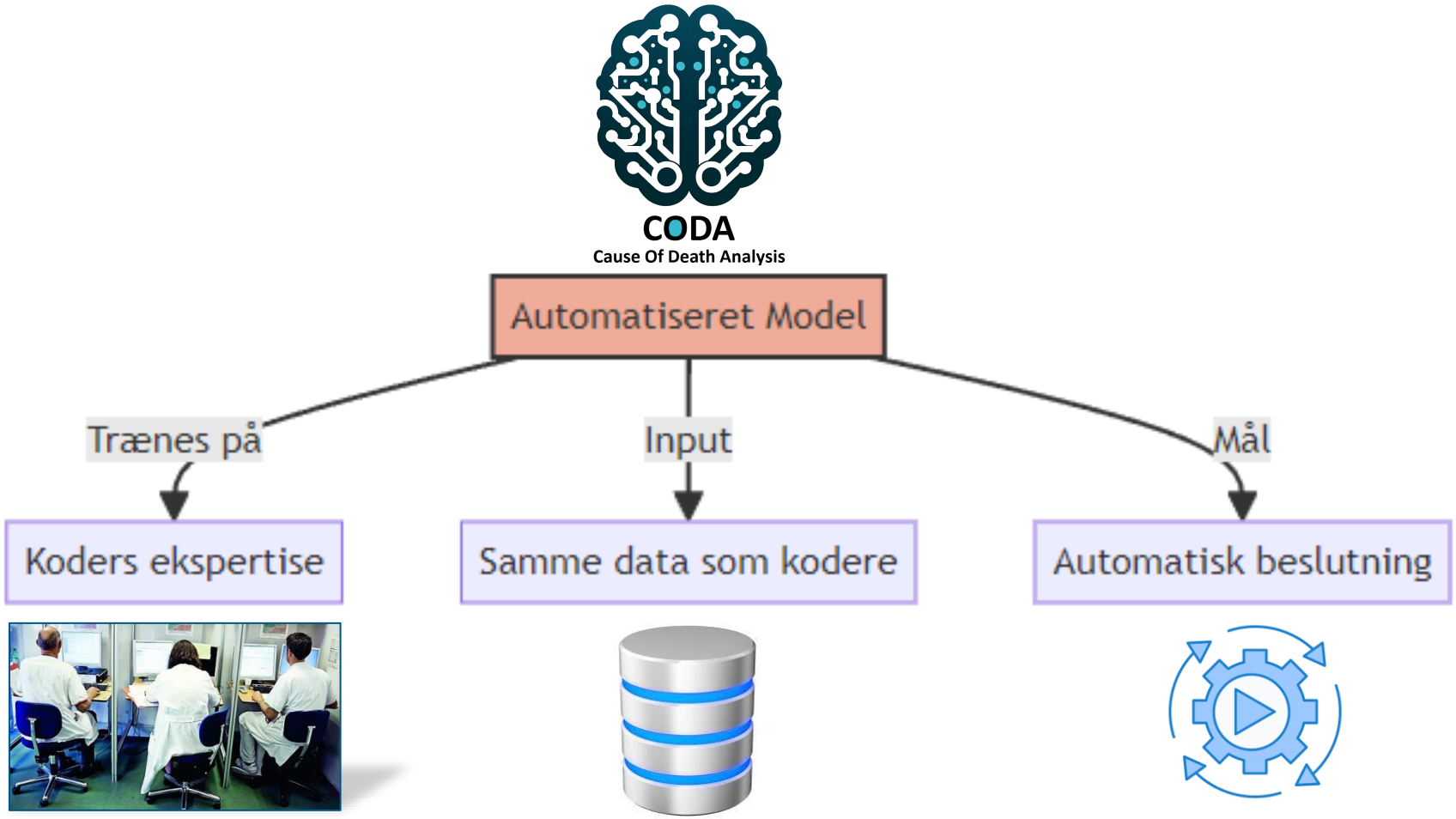


Proces for bestemmelse af den underliggende dødsårsag





Proces for bestemmelse af den underliggende dødsårsag



Proces for bestemmelse af den underliggende dødsårsag

Kodernes centrale rolle

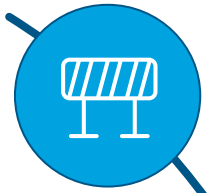
Eksperter

Træning af model

Validering af resultater

CODA
Cause Of Death Analysis

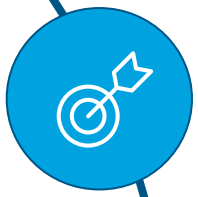
Klassifikation af underliggende dødsårsager i DAR



Problemstilling

Et datadrevet automatiseret beslutningsredskab til dødsårsager udestår

Det nuværende automatiske kodningsprogram IRIS kræver manuel intervention i 12-18% af indberettede dødsattester for at bestemme den underliggende dødsårsag. Der ønskes udviklet en ML-baseret metode til at effektivisere og understøtte denne manuelle proces



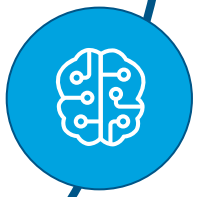
Mål

Udvikling af en automatiseret model til præcis bestemmelse af den underliggende dødsårsag i tilfælde, hvor IRIS-systemet viser sig utilstrækkeligt eller upålideligt

Data

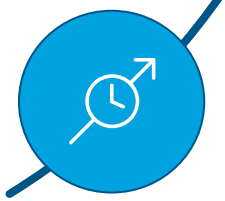
Datakilder for modellen:

1. Dødsårsagsregistret (DAR):
   - Dødsattestens side 1 og side 2
2. Landspatientregistret (LPR):
   - Patientkontakter, diagnoser, procedurer, forløbselementer, resultatindberetninger
3. Landsregistret for Patologi (PATO):
   - Diagnoser (SNOMED)
4. Oplysninger anvendt i den manuelle kodningsproces
   - Interviews (vægte)



Analyser

Multiklassifikationsalgoritme der kombinerer registerdata med maskinel analyse af evt. suppleret lægefaglig fritekst om dødsårsagen via en dansk-medicinsk sprog-model*



Mål og idéer

Algoritmen trænes på et datasæt bestående af tilfælde, hvor IRIS-systemet har givet usikre resultater. I disse tilfælde er den underliggende dødsårsag blevet manuelt fastlagt af fageksperter

DAR registeret opgøres med en forskydning på omkring et år. For eksempel, er data til og med 2022 tilgængelige i december 2023

*MeDa-BERT: A medical Danish pretrained transformer model (Pedersen et al., NoDaLiDa 2023)

SUNDHEDSDATA-STYRELSEN

Matrix på individniveau

Data til algoritme-træning:

- > 16.000 rækker (dødsattester) i et 80/20 split.
- > 12.000 kolonner med ophav fra: DAR, LPR, PATO og sprogmodel (fritekst)
  - Behandlede kategoriske og numeriske data fra registrene
  - Behandlede vektor-repræsentationer af semantisk forståelse (embeddings) fra sprogmodellen

Grov eksemplificering af opsætning:

CPR	DAR Side 1 Alder ved død	DAR Side 2 Antal diagnoser	LPR Antal kontakter med hovedspecialet i klinisk onkologi	IRIS-diagnose	Ekspert-diagnose	AI-diagnose	AI-diagnose-sandsynlighed fra modellen
123456-7891	71	10	1	W16	W18	W18	0,90
123456-7892	25	0	12	C619	C619	C618	0,20
123456-7893	76	3	15	D468	D468	D469	0,60
…	…	…	…	…	…	…	…

X_train/X_test y_train/y_test Model estimation

Matrix på individniveau

Data til algoritme-træning:

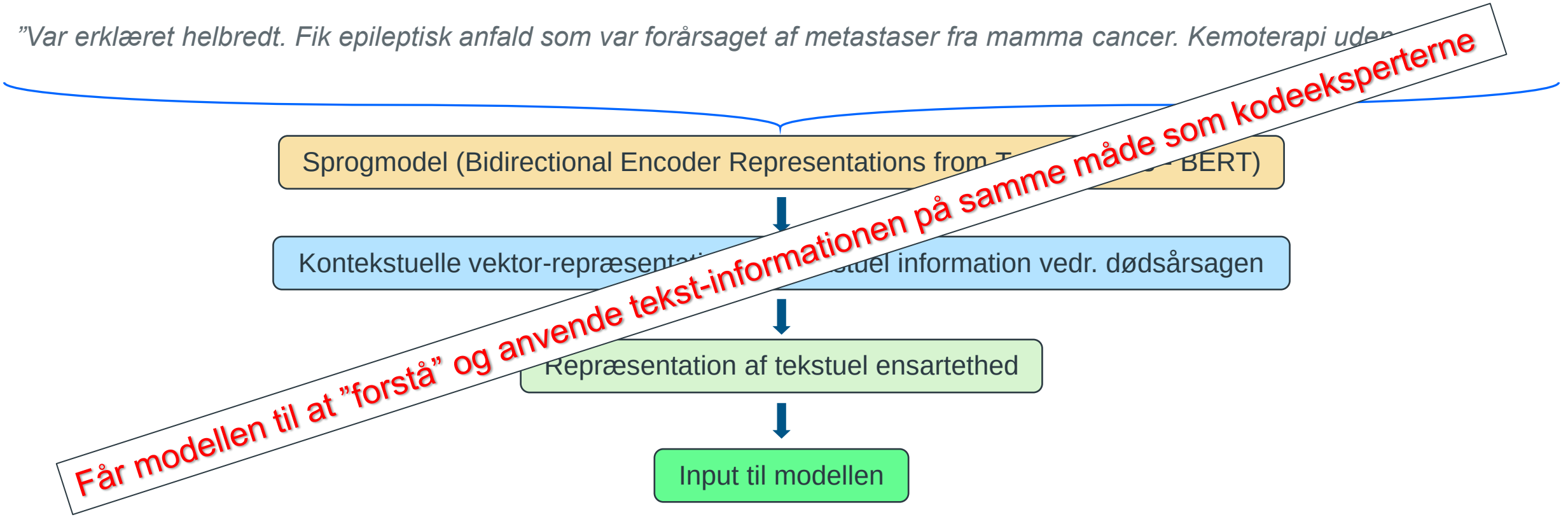
- > 16.000 rækker (dødsattester) i et 80/20 split.
- > 12.000 kolonner med ophav fra: DAR, LPR, PATO og sprogmodel (fritekst)
  - Behandlede kategoriske og numeriske data fra registrene
  - Behandlede vektor-repræsentationer af semantisk forståelse (embeddings) fra sprogmodellen

Sprogmodel

”Terminal lungekræft med lever- knogle - binyre samt hud-metastaser, i lindrende behandling. Indlægges til yderligere lindrende behandling.”

”Occult cancer med svær anæmi. Pt. ikke ønsket udredning. Sidste tid med pneumoni, så en konkurrerende dødsårsag.”

”Var erklæret helbredt. Fik epileptisk anfald som var forårsaget af metastaser fra mamma cancer. Kemoterapi uden



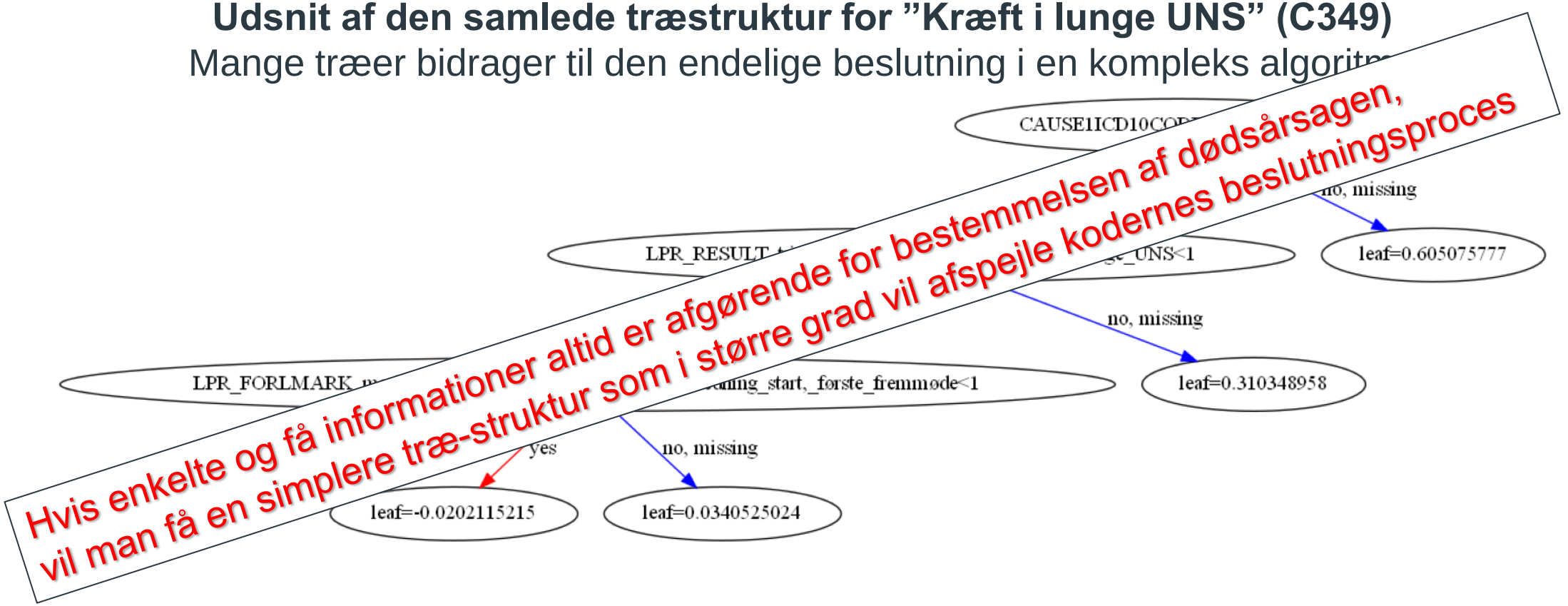
XGBoost

Træ-baseret model (”beslutningstræ”)

- Vigtigt at kunne gennemgående forstå og forklare modellens beslutninger
- En høj grad af gennemsigtighed gør det lettere at identificere og rette fejl

Udsnit af den samlede træstruktur for ”Kræft i lunge UNS” (C349)

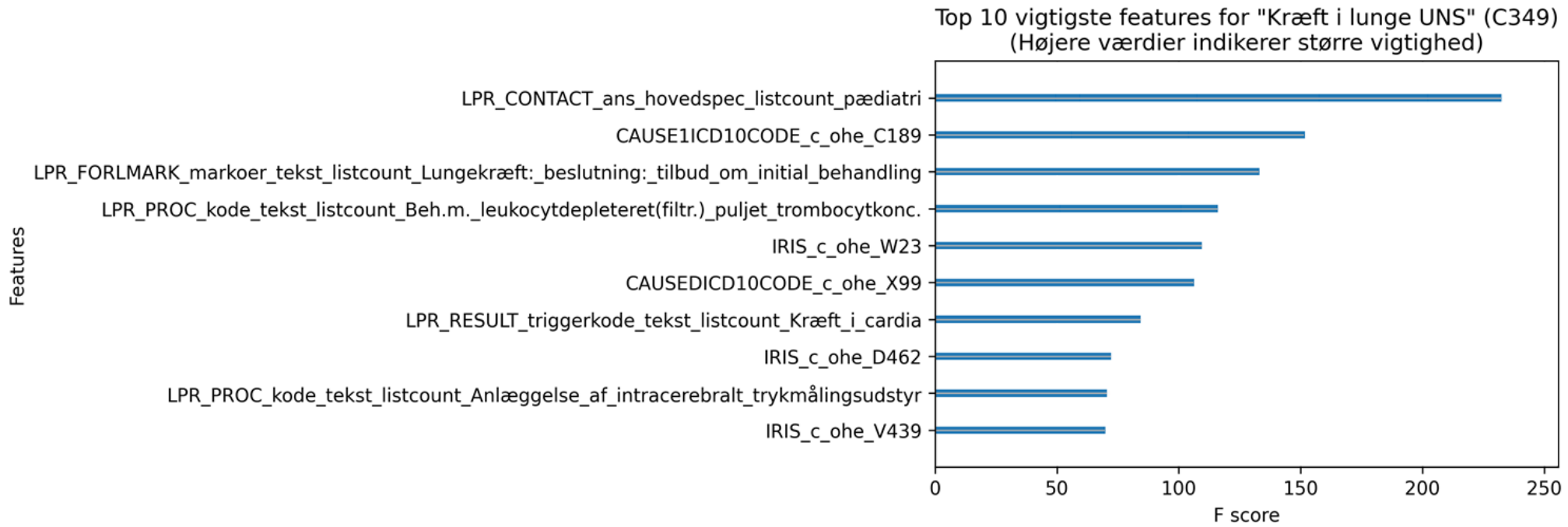
Mange træer bidrager til den endelige beslutning i en kompleks algoritme



Hvis enkelte og få informationer altid er afgørende for bestemmelsen af dødsårsagen, vil man få en simplere træ-struktur som i større grad vil afspejle kodernes beslutningsproces

XGBoost

Modellens feature importance (”forklaringselementer”)



Modelperformance

- For **72,2%** af 4202 attester i testdata var modellens angivne sandsynlighed på **>85%**

ICD10 klassifikation - som % af hovedtotal

Sum af freq		ai_sandsynlighed_prc_intv_no	ai_sandsynlighed_prc_intv_txt				
		1	2	3	4	5	Hovedtotal
AI_haandtering_ICD10	ai_match	<25	25-50	50-75	75-85	>85	
AI_handling	1	0,0	0,0	0,0	5,3	72,2	3.25
Man_kontrl_ej_match		2,3	3,1	2,6	0,0	0,0	3
Man_kontrl_match	1	0,8	2,5	6,2	0,0	0,0	0
Model_kontrl	0	0,0	0,0	0,0	1,0	4,0	211
Hovedtotal		3,1	5,6	8,9	6,3		4.202

- Af dødsattesterne med høj angivet modelsandsynlighed, var **94,8%** af modellens bud på dødsårsagen korrekt klassificeret ift. den manuelle håndtering

ICD10 klassifikation - som % af gruppetotal

Sum af freq		ai_sandsynlighed_prc_intv_no	ai_sandsynlighed_prc_intv_txt				
		1	2	3	4	5	Hovedtotal
AI_haandtering_ICD10	ai_match	<25	25-50	50-75	75-85	>85	
AI_handling	1	0,0	0,0	0,0	83,5	94,8	3.254
Man_kontrl_ej_match		74,8	55,1	29,6	0,0	0,0	337
Man_kontrl_match	1	25,2	44,9	70,4	0,0	0,0	0
Model_kontrl	0	0,0	0,0	0,0	16,5	5,2	211
Hovedtotal		100,0	100,0	100,0	100,0		4.202

Modelperformance

- For **72,2%** af 4202 attester i testdata var modellens angivne sandsynlighed på **>85%**

ICD10 klassifikation - som % af hovedtotal

Sum af freq		ai_sandsynlighed_prc_intv_no	ai_sandsynlighed_prc_intv_txt				
		1	2	3	4	5	Hovedtotal
AI_haandtering_ICD10	ai_match	<25	25-50	50-75	75-85	>85	
AI_handling	1	0,0	0,0	0,0	5,3	72,2	3.254
Man_kontrl_ej_match		2,3	3,1	2,6	0,0	0,0	337
Man_kontrl_match	1	0,8	2,5	6,2	0,0	0,0	400
Model_kontrl	0	0,0	0,0	0,0	1,0	4,0	211
Hovedtotal		3,1	5,6	8,9	6,3	76,1	4.202

- Af dødsattesterne med høj angivet modelsandsynlighed, var **94,8%** af modellens bud på dødsårsagen korrekt klassificeret ift. den manuelle håndtering

ICD10 klassifikation - som % af gruppetotal

Sum af freq		ai_sandsynlighed_prc_intv_no	ai_sandsynlighed_prc_intv_txt				
		1	2	3	4	5	Hovedtotal
AI_haandtering_ICD10	ai_match	<25	25-50	50-75	75-85	>85	
AI_handling	1	0,0	0,0	0,0	83,5	94,8	3.254
Man_kontrl_ej_match		74,8	55,1	29,6	0,0	0,0	337
Man_kontrl_match	1	25,2	44,9	70,4	0,0	5,2	400
Model_kontrl	0	0,0	0,0	0,0	16,5	100,0	211
Hovedtotal		100,0	100,0	100,0	100,0		

Usikkerhed i manuel kodning og modelforbedring

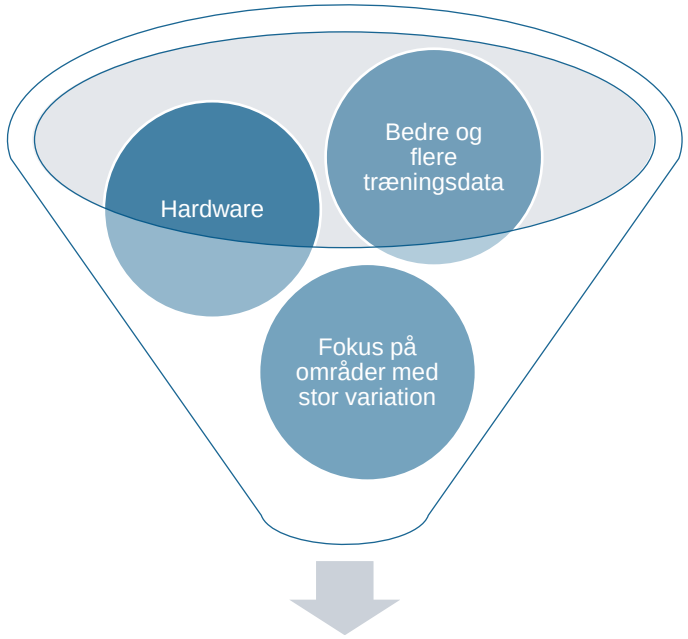
Afholdte valideringsøvelser (blindtest) viser betydelig variation i manuel kodning

- Usikkerheden på 5% ved høj modelsandsynlighed (>85%), er acceptabel sammenlignet med den menneskelige usikkerhed der er ved manuel kodning af den gruppe af dødsattester som de 5% repræsenterer (som er de særligt svære sager)

Variation i manuel kodning og modelforbedring

Afholdte valideringsøvelser (blindtest) viser betydelig variation i manuel kodning

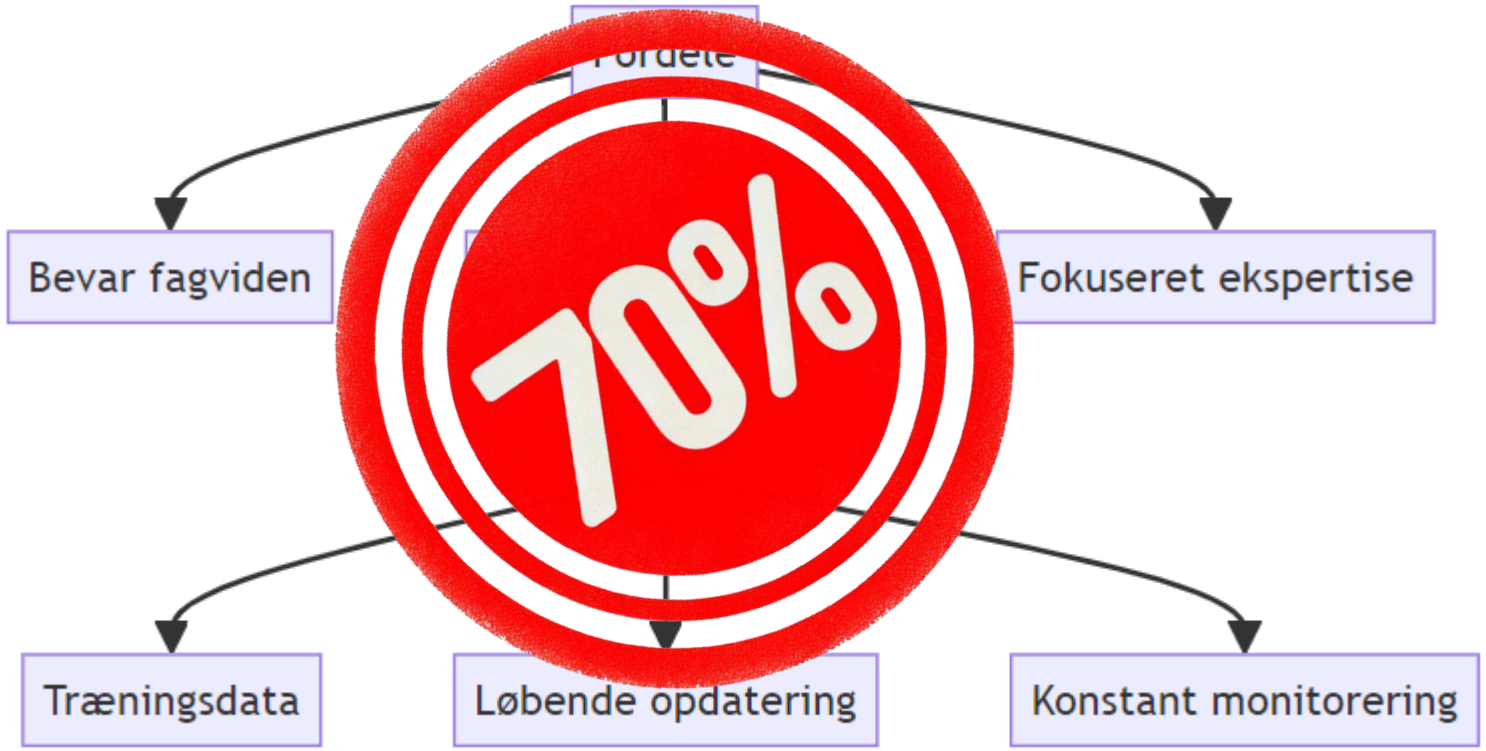
- De 5% forkert klassificerede dødsattester ved høj modelsandsynlighed (>85%), er acceptabelt sammenlignet med variationen i den manuelle kodning for den gruppe af dødsattester som de 5% repræsenterer (som er de særligt svære sager hvor koderne er usikre på dødsårsagen – det vil modellen også være). Modellen er, som koderne, usikker i klassifikationen af de komplicerede dødsattester.



Modelforbedring

Model kan fungere som en specialiseret medarbejder

Koderne var ikke umiddelbart imponerede over ML-modellen!



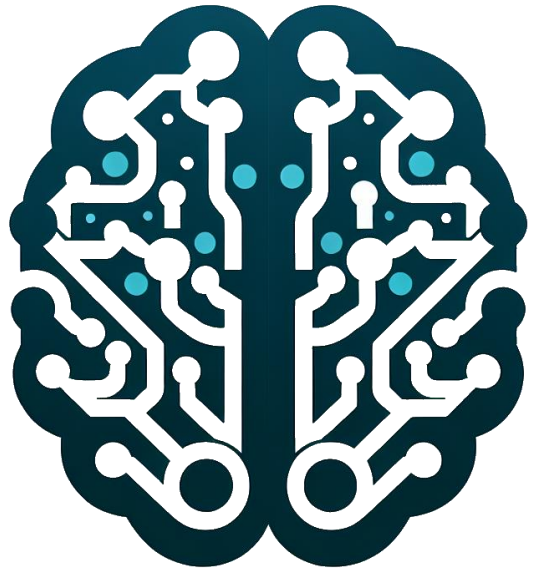
Udfordringer og observationer

- **Betydelig variation i manuel kodning, især ved øget kompleksitet**
- **Variation i manuel kodning påvirker modellens klassificeringsevne**
- **Lavfrekvente klasser i træningsdata påvirker modellens præstation negativt (håndteres)**
- **Begrænset forbedringspotentiale på modelsiden**
- **Fokus bør være på kvalitet/variation i ekspertvurderingerne (kvalitetsarbejde)**
- **Modelresultater kan udpege områder med størst forbedringspotentiale (målrettet kvalitetsarbejde)**
- **Modelresultater kan påvirke den manuelle kodning (håndteres)**

Hvor er SDS nu?

Undersøge mulighederne for anvendelse af modellen som beslutningsværktøj ved høje sandsynligheder

- Målrette kodeeksperternes indsat på at forbedre datakvaliteten – håndtering af særligt komplicerede dødsattester
- Mere målrettet og systematisk arbejde med integrationen af ML-modellen i arbejdet dødsårsagsregisteret forventes at forbedre datakvaliteten i registeret



CODA

Cause Of Death Analysis

Opsummering

Algoritmen kan bestemme den tilgrundliggende dødsårsag

- For **72,2%** af 4202 attester i testdata var modellens angivne sandsynlighed af sit valg på > **85%**
- Af disse, var **95%** af modellens bud på dødsårsagen korrekt klassificeret ift. den manuelle håndtering
- Resultaterne er kontrolleret ved omfattende valideringsøvelser med kodeeksperterne
- Usikkerheden på **5%** ved høj modelsandsynlighed, er acceptabel sammenlignet med den usikkerhed der er ved manuel kodning af den gruppe af dødsattester som de 5% repræsenterer (de særligt svære sager)

Algoritmen finder relevante features

- Flere forskellige machine-learning algoritmer er blevet trænet og testet
- XGBoost-algoritmen blev fundet bedst i denne case med en **nøjagtighed**[#] **på 0,85**
- Algoritmen finder relevante features (explainable AI)

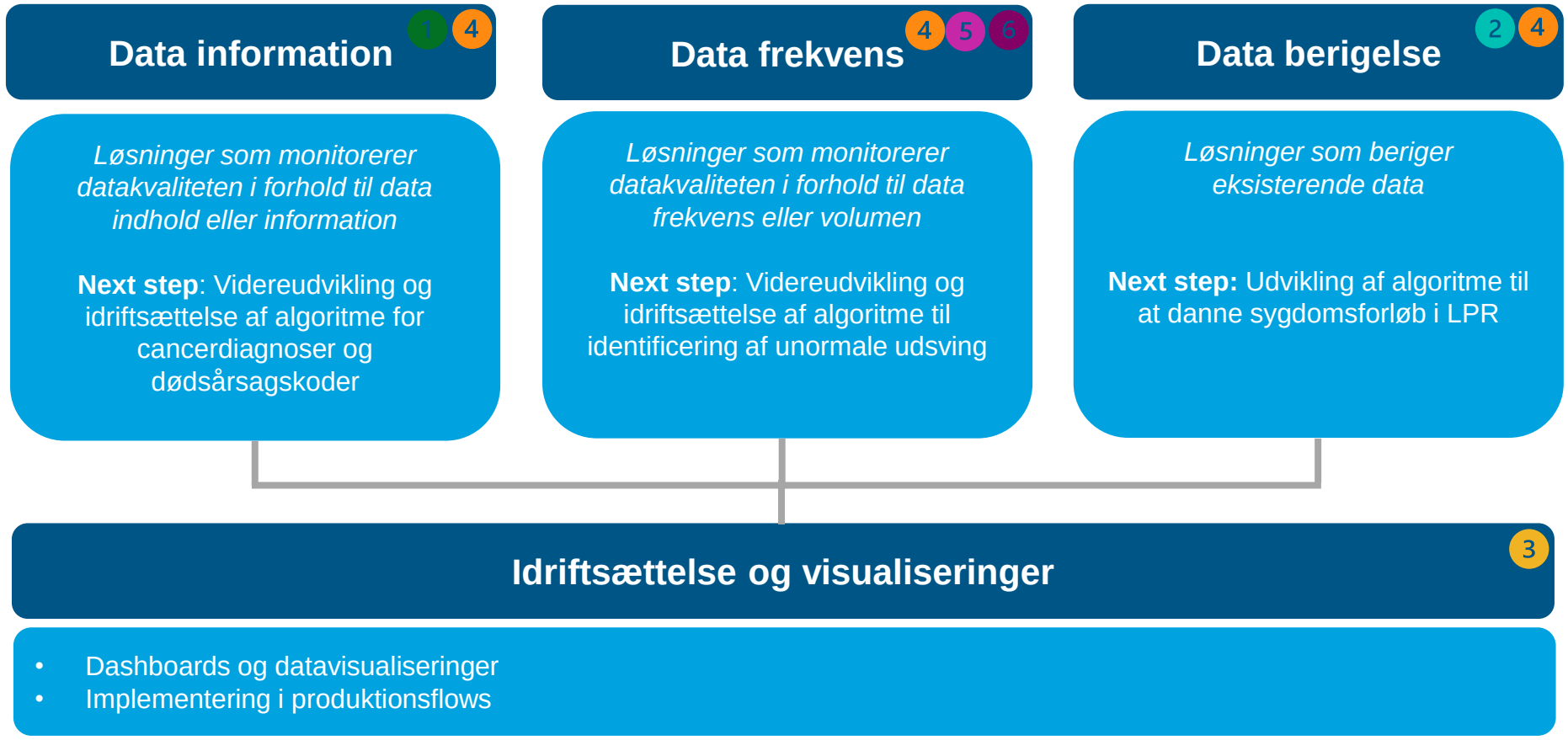
Algoritmen skal ibrugtages

- Formålet med at undersøge om det er muligt at bestemme den tilgrundliggende dødsårsag via en ML-model er påvist
- Algoritmen vil blive idriftsat som understøttelse til den manuelle ekspertkodning
- *Den ML-baserede løsning har potentiale til at blive et vigtigt redskab i det fremadrettede kvalitetsarbejde for dødsårsagsregisteret*

[#]Nøjagtighed er et udtryk for hvor nøjagtig eller pålidelig algoritmen er. Algoritmen er bedst, jo nærmere nøjagtigheden er på 1.

SUNDHEDSDATA-STYRELSEN

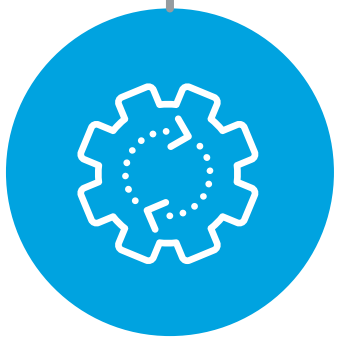
Idékatalog med kommende AI-baserede datakvalitetsprojekter i SDS



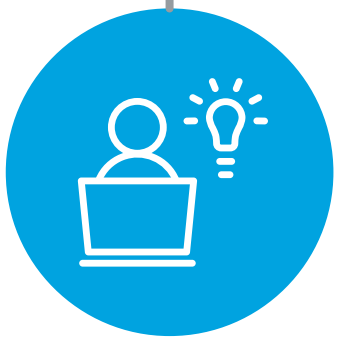
Foreløbige resultater og perspektiver for videre anvendelse

AI har et betydeligt potentiale i SDS og vil kunne anvendes på tværs af organisationen i kvalitetsarbejdet med de centrale sundhedsregistre

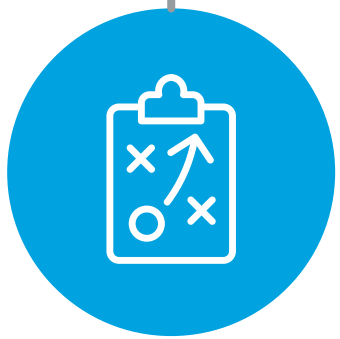
AI-metoder kan identificere væsentlige fejl og mangler i indberetninger - også for indberetningsområder hvor der er høj datakvalitet



AI kan automatisere tidskrævende og komplicerede datakvalitetsopgaver



Det er muligt at opbygge kompetencer for at anvende AI-baserede metoder til dataarbejdet i styrelsen



Det kræver målrettet arbejde og ressourcer at opbygge AI-kompetencer og tilhørende infrastruktur