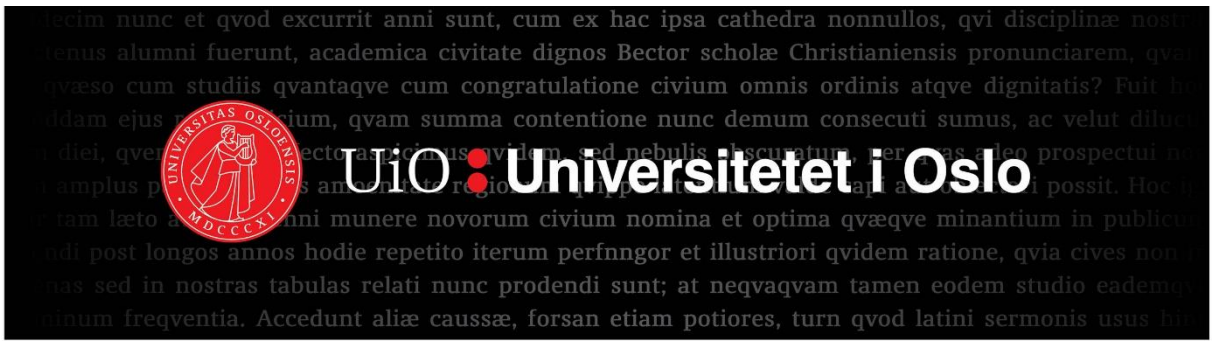

Interrater-reliabilitet for legegranskeres vurdering av helsehjelpsproblemer ved dødsfall i sykehus

En reliabilitetsstudie for journalgjennomgang med skjemaverktøyet PRISM2

Mette Utheim

Kandidatnummer 8

Mastergrad i interdisiplinær helseforskning

60 studiepoeng

Institutt for helse og samfunn

Medisinsk fakultet

Dato 15/5 2022

Forord

Min interesse for pasientsikkerhet oppstod da jeg skrev fordypningsoppgave i nyfødtsykepleie om dobbeltkontroll av legemidler, avviksmeldinger og pasientsikkerhetskultur. Da jeg noen år senere lette etter et tema for masteroppgaven i interdisiplinær helseforskning ble jeg tatt entusiastisk imot på avdeling for pasientsikkerhet, kvalitet og samhandling ved OUS, og fikk tilgang til rådata fra deres pågående journalgjennomgangsstudie.

Gjennom arbeidet med masteroppgaven har jeg fått bedre forståelse for begrepene validitet og reliabilitet innen måleteorien. Jeg har reflektert over hvordan utfallsrommet som defineres av variablene og designaspekter ved reliabilitetsstudien påvirker informasjonen journalgjennomgangsmetoden produserer, og har ikke minst blitt overrasket over hvor vanskelig det kan være å måle fenomener vi i utgangspunktet vet mye om. Jeg har blitt kjent med hvordan variablenes målenivå kan påvirke datafordeling og enighet, og hvordan dette igjen kan påvirke mye brukte statistiske mål for interrater- reliabilitet. Det største læringsutbyttet har vært forståelsen av hvordan valg av design og grad av rigiditet i utførelsen kan påvirke variasjonen i målingene, og hvordan valg av statistisk metode kan påvirke tolkningen av resultatene. Annerkjennelsen av det ukjente antallet subjekter som er vanskelige å skåre og vil kamufleres av designet der en har ett måletidspunkt eller èn gransker, men like er fullt til stede og påvirker variasjon i skårene og dermed reproduserbarhet, ser jeg som svært verdifull. Veien til innlevering har vært lang og lærerik, også når det gjelder studieteknikk og prioriteringer.

Tusen takk for inkludering, all tålmodighet og veiledning til Dr. Anne Karin Lindahl og Dr. Trine Sand Kaastad.

Takk til Thomas Jørgensen Riiser ved Avdeling for kvalitet, virksomhets- og risikostyring for hjelp med tekniske løsninger og tilrettelegging under pandemien.

Takk til Luigi Maglanoc ved IT- support UIO for å ha ryddet opp i alle feilmeldingene i R- Studio, og Reza Ghiasvand ved OUS for hjelp med statistikkforståelsen.

Til gulla mine, Leif, Johann, Sverre og Oliver,

- nå skal jeg ikke studere mer.

Sammendrag

Design: Reliabilitetsstudie **Hensikt:** Å undersøke interbedømmer- reliabilitet for hovedkonklusjonene fra legegranskeres journalgjennomgang med PRISM2 skjemaet i norsk oversetting. **Bakgrunn:** Uønskede hendelser og unngåelige dødsfall påfører pasienter og pårørende unødig lidelse, svekker tilliten til helsetjenestene i befolkningen, og utgjør en stor helseøkonomisk utgiftspost. Journalgjennomgang er den mest brukte metoden for identifikasjon av helsehjelpsproblmer, men er funnet å produsere lav til moderat interrater-reliabilitet (IRR). Oppgaven besvarer delmål en studie ved avdeling for pasientsikkerhet og samhandling ved Oslo Universitetssykehus (OUS), og har problemstillingen: *«Hvordan påvirkes vurderinger av interrater- reliabilitet for hovedkonklusjonene fra journalgjennomgang av to sammenliknbare reliabilitetsmål, og hva er forekomsten av helsehjelpsproblemer i vårt utvalg for disse gjennomgangene?»* **Teori:** Måleteori, bedømmelsesteori og pasientsikkerhetsteori ble drøftet, med fokus på validitet, reliabilitet, tilfeldig enighet, bekreftelsesfeller for ekspertgranskeren og tendenser for de utvalgte IRR- målene. **Metode:** Uavhengige journalgjennomganger ble utført av to legegranskere for et tilfeldig utvalg journaler (*n*=200, Kvinner *n*=93) fra alle dødsfall ved OUS i 2014 (N=1081). Interrater- reliabilitet ble beregnet for hovedkonklusjonene ved Prosent enighet, Cohens Kappa og Gwets AC1. **Resultater: V**ed enighet mellom granskerne ble det påvist et visst belegg for helsehjelpsproblemer i 33% (*n*= 66, kvinner *n*=15) av journalene. Prosent enighet for helsehjelpsproblemer var 86.2%., Kappa .268 (KI .128 - .406 *P*- verdi <.001) og AC1 .384 (KI .209 - .488, P- verdi <.05). Det ble konkludert med over 50% sannsynlighet for unngåelig dødsfall i 2% av journalene (*n*=4, kvinner *n*=1). Prosent enighet for unngåelig dødsfall var 88.5%, Kappa .209 (KI -.006-.424, P- verdi <.001), Gwets AC1 .865 (KI .806 - .925, P- verdi <.001). **Konklusjon:** Forekomsten av helsehjelpsproblemer samsvarer med tidligere sammenliknbare studier. AC1 var høyere enn Kappa for alle konklusjoner, og ga forskjellige kvalitative tolkninger for unngåelig dødsfall. Forskjellene varierte med datafordelingen og kan være forårsaket av kappaparadoksene. **Begrensninger:** Det er ikke undersøkt om pasientene hadde overlevd sykehusoppholdet uten helsehjelpsproblemer. Utvalgsvaliditeten er ikke tilfredsstillende med tanke på å undersøke helsehjelpsproblemer som ikke fører til dødsfall. Journalgjennomgang har velkjente validitets- og reliabilitetsutfordringer. Kappa er konservativ og AC1 er liberal ved skjeve fordelinger, og effekten av dette er ukjent. Dikotomisering ga informasjonstap med en dobling av forekomsten av unngåelig dødsfall samt økning av reliabilitet målt ved Prosent enighet, og metodevalget begrenser verdien av funnene for validering av PRISM2- skjemaet.

Abstract

Design: Reliability study **Purpose**: To examine the inter-rater reliability (IRR) of the main conclusions from Retrospective Case Record Review (RCRR) with the Norwegian translation of the PRISM2 instrument. **Background:** Adverse events and preventable deaths inflict unnecessary suffering on patients and their next of kin, challenges the trust the healthcare services enjoy with the larger public, and are associated with large financial costs. Retrospective case record review (RCRR) is the preferred method for detection of problems of healthcare but produce poor to moderate Inter-rater reliability (IRR). This thesis answers aims included in a study at the department of patient safety and coordination at the Oslo University Hospital (OUS) with the research question: *"How are assessments of inter- rater reliability for the main conclusions from retrospective case record review affected by two comparable reliability indices, and what is the prevalence of problems in care in our sample,as identified by the review?"* **Theory**: Measurement theory as well as heuristics and patient safety theory were considered with a focus on validity, reliability, agreement by chance, biases of the expert rater and performance of interrater- reliability indices. **Method:** Independent RCRRs were conducted by two medical expert raters on a random sample (*n*=200, Women n=93) from in- patient deaths at OUS in 2014 (N=1081). Interrater reliability (IRR) was by percent agreement, as well as the chance- corrected indices Cohens Kappa and Gwets AC1. **Results:** Agreement between raters determined the prevalence of some evidence of problems of healthcare within the case notes to be 33% (*n*= 66, women *n*=15). Percent agreement was 86.2%., Kappa .268 (CI.128 - .406 *P*- value <.001) andAC1 .384 (CI.209 - .488, P- value <.05). Avoidability of deaths with a 50% likelihood had a prevalence of 2% (*n*=4). Women comprised 25% (*n*=1). Percent agreement was 88.5%, Kappa .209 (CI -.006-.424, P- value <.001), Gwets AC1 .865 (CI .806 - .925, P- value<.001). **Conclusion:** Prevalences of problems of care were comparable to that of previous studies.AC1 was higher than Kappa for all findings, og the indices gave different qualitative interpretations (Fair / Almost perfect) for the avoidability of deaths. The differences in IRR varied with prevalence and may be affected by the Kappa paradoxes. **Limitations:** It is not known if the patients would have survived the hospitalization without problems of healthcare, and the sampling method dismisses detection of problems in care in the majority of patients discharged alive. RCRR has well- known challenges related to validity. Kappa is a conservative, and AC1 is a liberal indice when applied to skewed distributions. Collapsing variables led to a loss of information, doubling of the prevalence of avoidable deaths (*n*=2 to *n*=4), and improvement of reliability measured by Percent agreement. This choice of method limits the value of the reliability- evaluation as it relates to validation of the PRISM2- instrument.

Keywords: IRR, Inter-rater(/coder/examiner/observer/judge) reliability/agreement/consensus, Measures of concordance, Medical Expert Rater, Expert heuristics, Retrospective Case note/ Case Record/ Chart review, RCRR, Adverse events, Problems in Health care, Preventable deaths, Ameniable deaths, Avoidable deaths, Medical error, In-hospital deaths, In-patient deaths, Quality of care, Indices of patient safety PRISM2, Preventable Incidents, Survival and Mortality Study,

Tabeller

Tabell 1 Enighet (Uthevet diagonal) med total observert enighet (Nedre høyre celle) 32
Tabell 2 Uthevet uenighet 33
Tabell 3 Granskernes marginaling og utregning av marginalfordeling 36
Tabell 4 Utregning av prosent enighet 37
Tabell 5 Gwets AC1 utregning 40
Tabell 6 Felles paradokser for Cohens Kappa og Gwets AC1 42
Tabell 7 Særegne paradokser ved Cohens Kappa (46) 43
Tabell 8 Gwet AC1s særegne paradokser 45
Tabell 9 Kvalitative tolkninger av IRR 48
Tabell 10 Egenskaper ved legegranskerne 50
Tabell 11 Variabelutvalg med omforming 52
Tabell 12 Endring i enighet og foredling som følge av dikotomisering 54
Tabell 13 WHO aldersgrupper og sentralitetsmål 60
Tabell 14 Prosent enighet, Cohens Kappa og Gwets AC1 for hovedkonklusjonene 61
Tabell 15 Helsehjelpsproblemer 62
Tabell 16 Unngåelig dødsfall 63
Tabell 17 Helsehjelpsproblemer Kvinner 64
Tabell 18 Unngåelig dødsfall Kvinner 65
Tabell 19 IRR hovedkonklusjoner Kvinner 65

Figurer

Figur 1 Høy assosiasjon og lav enighet eksempel 34
Figur 2 Utregning av tilfeldig enighet basert på antall kategorier 36
Figur 3 Cohens Kappa utrenging 39
Figur 4 Enighetstabell unngåelig dødsfall (ordinal) 62

Vedlegg 1 Personvernombudets tilråding 89
Vedlegg 2 Studieprotokoll 91
Vedlegg 3 PRISM2 Skjema for journalgjennomgang 98
Vedlegg 4 Antagelser og paradokser fra Zhao, Liu og Deng (2013) (forenklet). 108
Vedlegg 7 Øvrige analyser 109

Innhold

1.0 Bakgrunn 8
1.2 Hensikt og målsetning 9
1.3 Problemstilling og forskningsspørsmål 10
1.4 Fokus og begrensning 10
1.5 Oppbygning 10
1.6 Begrepsbruk 11
1.4 PRISM2 12
4.0 Teoridel 13
2.1 Helsehjelpsproblemer 14
2.2 Unngåelige dødsfall 16
2.3 Identifikasjonsmetoder 17
2.3.1 Overdødelighet 17
2.3.2 Elektroniske avvikssystemer 18
2.3.3 Dødsattester, obduksjon og melding om mulig unaturlig dødsfall. 18
2.3.4 Journalgjennomgang 19
2.4 Bedømmelsesteori og bekreftelsesfeil for ekspergranskeren 22
2.5 Forskning på helsehjelpsproblemer med reliabilitetsberegning 24
3.0 Måleteori 30
3.1 Validitet 30
3.2 Validitet som spesifisitet og sensitivitet 30
3.3 Reliabilitet, enighet og uenighet 31
3.4 Assosiasjon og intern konsistens 33
3.5 Inter- rater reliabilitet og enighet 34
3.6 Tilfeldig enighet 35
4.0 Beregning av Interrater- reliabilitet 37
4.1 Prosent enighet 37
4.2 Korrigering for tilfeldig enighet 38
4.2.1 Tilfeldighetskorrigerte mål: Cohens Kappa 39
4.2.2 Tilfeldighetskorrigerte mål: Gwets AC1 40
4.2.3 Antagelser for Cohens Kappa og Gwets AC1 41
4.2.4 Paradokser 42
4.2.5 Felles paradokser for Cohens Kappa og Gwets AC 42
4.2.6 Særegne paradokser for Cohens Kappa 43
4.2.7 Særegne paradokser for Gwets AC1 45
4.3.8 Generelle observasjoner 46
4.3 Tolkning av IRR-koeffisienten 47
5.0 Metode 49

5.1 Design 49
5.2 Utvalg 49
5.2 Granskerutvalg 50
5.3 Datainnsamling 50
5.3.1 Variabelutvalg 51
5.3.3 Omforming av variabler 53
5.3.4 Manglende data 55
5.4 Statistiske metoder og presentasjonsvalg 57
Ekskluderte koeffisienter 58
6.0 Veiledere og ressurser 59
6.2 Etiske hensyn og personvern 59
6.0 Resultater 59
6.1 *Alder* og *Kjønn* i utvalget 60
6.2 Oppsummering Interrater- reliabilitet 60
6.3 Helsehjelpsproblemer 61
6.4 Unngåelige dødsfall- ordinal variabel 62
6.5 Unngåelig dødsfall- dikotomisert 62
6.6 Hovedkonklusjoner Menn 63
6.7 Hovedkonklusjoner Kvinner 64
7.0 Diskusjon 65
8.0 Konklusjoner 80
8.1 Begrensninger og anbefalninger for videre forskning 81
9.0 Litteraturliste 84

1.0 Bakgrunn

Den mye omtalte Amerikanske rapporten *To err is human*, antyder at flere amerikanere dør av medisinske feil enn av brystkreft, og at kostnadene av helsehjelpsproblemer kan være opptil 29 milliarder USD årlig (1). Internasjonale forskningsoppsummeringer har funnet at uønskede hendelser kan ramme mellom 2,9% og 8% av sykehuspasientene, og at så mye som 8.6 % av hendelsene kan ha dødelig utfall(2, 3).

Siden 2010 blir alle offentlige sykehus i Norge overvåket med selvrapporteringsverktøy[1], og forekomsten av helsehjelpsproblemer ble i årene 2010- 2013 med denne metoden funnet å være mellom 16,1% og 13% (4). En journalgjennomgang fra 2018 for sykehusdødsfall ved St. Olavs hospital med et videreutviklet skjemaverktøy fra PRISM- studiene fant at mellom 4,2% og 6.7% av dødsfallene var unngåelige, avhengig av bevisbyrde(5), og ved det samme sykehuset ble det i 2011 funnet en forekomst av unngåelige dødsfall på 2.9%. ved granskning med et legepanel og en mortalitetsindeks- instrument(6)

Veilederen til *Forskrift om ledelse og kvalitetsforbedring i helsetjenesten* fastsetter at helsetjenestene skal være virkningsfulle, trygge, samordnede og involvere brukere, samt utnytte ressursene (7, 8). Avdelingsdirektør ved kunnskapssenteret i folkehelseinstituttet, Dr. Anne Karin Lindahl, og seniorforsker Lise Lund Håheim, sammenfatter i artikkelen *Helsetjenesteforskning og helsetjenestens kvalitet* fra 2017 helsetjenesteforskningens hensikt som å belyse om slike kvalitetskriterier oppfylles, og hevder at forskningen kun er nyttig om den lykkes i å beskrive praksisvariasjonen(9). Lindahl og Håheim presenterer utfordringene helsetjenesteforskningen står ovenfor sett fra fagmiljøenes ståsted, og fremhever behovet for bredere metodekompetanse som en begrensning. Forfatterne påpeker at enighet om kriteriene for datakvalitet er avgjørende for Norges bidrag i systematisk sammenligning av kvaliteten på helsetjenestene internasjonalt, såkalt *Benchmarking(9)*.

Dr. Trine Sand Kaastad, som leder Avdeling for pasientsikkerhet og kontinuerlig forbedring ved Oslo universitetssykehus, tilføyer at pasientkommunikasjonen ved slike gjennomganger følges mellom mange aktører og over lang tid, og at det er avgjørende at verktøyene som benyttes er validerte og at det er godt samsvar mellom vurderingene (Kaastad, Trine S. Re: Veiledning masterstudenten Email: tkaastad@ous-hf.no, 24.09.2021). Kaastad bemerker at det per idag ikke finnes validerte verktøy for vurdering av helsehjelpsproblemer ved journalgjennomgang, og at datakvaliteten påvirkes av at fagmiljøene må benytte indirekte metoder (Ibid). Slike metoder har

[1] Global trigger tool (GTT)

demonstrert validitetsutfordringer, eksemplifisert ved en gjennomgang av dødsattester og innrapportering av meldepliktige dødsfall av Alfsen et. al fra AHUS i 2013 (10). Her fant forfatterne fant store mangler i dokumentasjonsgrunnlaget, foruten manglende innrapportering av sannsynlig meldepliktige dødsfall. Alfsen et al konkluderte med at gjennomgang av sykehistoren ved død bør være fast rutine og håndteres av en definert legegruppe(10).

Det formidles av flere forfattere at journalgjennomgang er et svært ressurskrevende arbeid, som må forsvares med at resultatene er pålitelige(3, 9, 11, 12). Utvikling av egnede overvåkningsverktøy er et satsningsområde for forbedringskunnskapen i helseforetakenes pasientsikkerhetsarbeid(4), og i 2006 ga norsk pasientregister ut en håndbok i journalgjennomgang, for å bedre datakvaliteten i sykehussektorens kvalitetsarbeid(11).

Ved avdeling for pasientsikkerhet og samhandling ble det i 2015 påbegynt en studie ved OUS ved navn «Pasientdødsfall i Oslo Universitetssykehus vurdert ved retrospektiv journalgjennomgang (RCRR)», der dødsfall i OUS fra 2014 (N=1081) undersøkes med forskjellige metoder for deteksjon av helsehjelpsproblemer for å tallfeste forekomsten av disse og sammenlikne to mye brukte skjemaverktøy for journalgjennomgang (Vedlegg 2). Denne masteroppgaven inngår i en pilotstudie med uavhengige journalgjennomganger av et utvalg (n=200) for å vurdere reliabilitetsaspekter og brukervennlighet ved journalgjennomgangsskjemaet PRISM2.

1.2 Hensikt og målsetning

Oppgavens hensikt er å bedre beslutningsgrunnlaget for valg av instrument for hovedstudien og sammenfatte metodologiske aspekter ved interrater- reliabilitet for fagmiljøet som arbeider med pasientsikkerhet ved OUS.

Oppgaven går under punkt 1 og 3 i den pågående studien "*Pasientdødsfall i Oslo Universitetssykehus vurdert ved retrospektiv journalgjennomgang (RCRR)"* (Vedlegg 2), og har følgene målsetninger:

1: Å beregne interrater- reliabilitet for hovedkonklusjonene fra journalgjennomgangene ved prosent enighet, Cohens Kappa og Gwets AC, for hele utvalget samt aldersgrupper og kjønn.

2: Å drøfte innvirkningen av datafordeling på disse målene samt mulige årsaker til forskjeller i skår

1.3 Problemstilling og forskningsspørsmål

«Hvordan kan forskjeller i forskjellige mål for interrater- reliabilitet ved legegranskeres journalgjennomgang med skjemaverktøyet PRISM2 forklares, og hva er forekomsten av helsehjelpsproblemer i vårt utvalg fra sykehusdødsfall i 2014?»

1. Hva er forekomsten av uønskede hendelser og forebyggbare dødsfall for utvalget, og i hvor mange journaler er det enighet for disse vurderingene?

2. Hvordan kan tendenser for variasjon i skår mellom granskerne forklares?

3. Hva er interrater- reliabilitet målt ved Prosent enighet, Cohens Kappa og Gwets AC1 for variabelutvalget

4. Kan forskjeller mellom målene forklares av deres følsomhet for datafordelingen?

1.4 Fokus og begrensning

Oppgavens fokus er reliabilitetsaspekter ved journalgjennomgang for deteksjon av helsehjelpsproblemer.

Litteratur og statistiske mål for IRR som vurderes begrenses til et scenario for reliabilitetsstudien med to granskere og kategoriske datatyper med dikotomt målenivå.

Gwets AC1 er en hittil lite brukt koeffisient for journalgjennomgang og at litteraturgjennomgangen fant ingen relevante studier med journalgjennomgang der AC1 er brukt, vurderingen av denne baserer seg derfor i stor grad på to metodeartikler samt Gwet (2008).

1.5 Oppbygning

I de innledende kapitlene er bakgrunn, problemstillingen med forskningsspørsmål, fokus for og begrensing av oppgavens tematikk presentert.

Etter avklaring av begrepsbruk følger en oppsummering av egenskaper ved skjemaverktøyet PRISM2 og deretter en definisjoner av helsehjelpsproblemer og unngåelige dødsfall samt en oppsummering av de mest brukte identifikasjonsmetodene for disse.

Så presenteres bekreftelsesfeller som rammer ekspertgranskeren hentet fra bedømmelsesteori, før relevant forskning med journalgjennomgang som identifikasjonsmetode for helsehjelpsproblemer gjennomgås.

Teorikapitlet innledes med sentrale begreper fra måleteorien og disse relateres til reliabilitetsdesignet. Så beskrives egenskaper ved de utvalgte målene for interrater- reliabilitet, med antagelser og paradokser samt forskjellige tolkninger av koeffisientene.

Metodekapitlet gir avklaring av beskrivelser av design, subjekt- og granskerutvalg og valg gjort for datasammenfatningen og variabelutvalget, med vekt på kollapsing/ dikotomisering av variabler og sider ved variablenes semantiske validitet som kan påvirke reliabilitetsberegningen.Metodekapitlet avsluttes med begrunnelser for valg av statistisk metoder og presentasjonsmetoder, samt resursser, etiske hensyn og persovernhensyn for oppgaven.

I resultatkapitlet presenteres først alder og kjønnsrepresentasjon i utvalget før interrater-reliabilitetsberegningene oppsummeres. Deretter beskrives resultatene fra disse beregningene i sammenheng med enighetstabeller for de to hovedkonklusjonene helsehjelpsproblem og unngåelig dødsfall, for utvalget som helhet og for kvinner og menn i utvalget. Øvrige analyser som ikke angår hovedkonklusjonene legges ved (vedlegg 6).

Avslutningsvis kontekstualiseres resultatene med gjennomgått teori og forskning, og til slutt presenteres konklusjoner med begrensninger for overførbarhet av resultatene og anbefalninger for videre forskning.

1.6 Begrepsbruk

Med enkelte unntak, er det tilstrebet å anvende norske ord og uttrykk. Der det er gjort spesielle valg i oversettelsen utover begrepsavklaringen, avklares dette i fotnote.

Chance oversettes *tilfeldig*, ettersom *sjan*se på norsk oppleves å ha et noe annet meningsinnhold i retning «forsøk» og «risiko».

Unngåelige dødsfall erstatter engelske *preventable, avoidable* eller *amenable deaths/ mortality*, og er det begrepet som er valgt for den norske oversettelsen av PRISM2. *Forebyggbare dødsfall* anses å være et synonym for vår oppgave.

Gransker er ment å dekke engelske begreper som «Coder» «Rater» «Judge», «observer» og «examiner». Hvilket begrep som brukes begrunnes sjelden i forskningen, men Bauer (2000) beskriver «Coder», som et kvalitativt begrep(13). Det oppleves at *Gransker* er dekkende for ekspertkompetanse og skjønnsmessige vurderinger.

Helsehjelpsproblem er et begrep introdusert av Hogan (2014), som erstattet uønskede hendelser og er en forutsetning for unngåelig dødsfall. *Helsehjelpsproblemer* brukes også som samlebetegnelse i betydningen *uønskede hendelser og unngåelige dødsfall*

Interrater- reliabilitet: Det norske begrepet *interbedømmerreliabilitet* oppleves noe omstendelig, og det engelske uttrykket er godt innarbeidet på norsk.

Marginalfordeling dekker på engelsk både *marginal distribution* og *quota og* beskriver den relative fordelingen av skår, avlest i marginalene i en enighetstabell.

Observert enighet er i denne oppgaven det *n* tall som avleses i diagonalen i en enighetstabell, mens *enighet* brukes i den allmenne betydningen av fenomenet enighet.

Oppriktig skåring erstatter engelske *honest,* som i litteraturen har betydningen *velinformert*, uten at det antyder at skåringer med usikkerhet ikke kan være oppriktige.

Skåring erstatter engelske ord som *coding, making assignments* og *making categorisations,* selv om en ser engelsk *score* oftest brukt for kontinuerlige variabler. Koding velges bort etter samme begrunnelse som at *gransker* velges for *coder,* og *kategorisering* oppleves språklig tungt ettersom granskeren kategoriserer i kategorier.

Bekreftelsesfeller og Bekreftelsesfeil erstatter engelske *bias* der temaet er bedømmelsesteori. Der *Bias* beskriver skjevhet i utvalg brukes *utvalgsskjevhet*

Det benyttes forkortelser utover i oppgaven, som OUS, RCRR, IRR, Kappa og AC1. Første gang ordene brukes skrives de ut.

1.4 PRISM2

Oppgaven bygger på data fra uavhengige journalgjennomgjennomganger utført av legegranskere med skjemaverktøyet som ble utviklet for standardisering av journalgjennomgangene i Preventable Incidents, Survival and Mortality Study 1 og 2, eller PRISM-studiene av Hogan et al. i 2014 og 2015(14, 15). Skjemaet ble oversatt til norsk etter retningslinjer for oversetting ved Avdeling for Pasientsikkerhet og Samhandling ved OUS (Vedlegg 2).

Fokus for journalgjennomganger som struktureres etter PRISM2 er gjeldende sykehusinnleggelse og de helsehjelpsproblemer som var direkte assosiert med dødsfallet(16). Skjemaet instruerer granskerne i å gjennomgå alle dokumenter i pasientjournalen kronologisk, som henvisninger,

laboratorieresultater og bildediagnostikk, samt alle lege- og sykepleienotater (Vedlegg 3, 14). Om det foreligger obduksjonsrapport eller dødsattest leses disse til sist for å unngå bekreftelsesfeil som følge av etterpåklokskap(16).

For å identifisere forebyggbare dødsfall og uønskede hendelser, skilles det mellom komplikasjoner som oppstår *på tross av* at helsehjelpen har vært av akseptabel standard, og *problemer i helsehjelpen,* definert som *enhver hendelse som har inntruffet som følge av at helsehjelpen ikke har møtt faglige standarder og har ført til helseskade (16).* En slik hendelse kan være manglende inngripen, som en manglende diagnose eller ikke iverksatt behandling, eller aktive handlinger[2], som uriktig behandling eller organisering av helsehjelpen (12).

Granskerne instrueres i å gjøre skjønnsmessige vurderinger av hvordan helsehjelpen ideelt sett skulle vært organisert, etter hvilke akseptable faglige standarder og normer som gjelder for det enkelte tilfellet(16). For vurderingen av unngåelig dødsfall skal granskeren spørre seg om dødsfallet var forventet, eller om det inntraff som følge av kvaliteten på helsehjelpen heller enn den naturlige progresjonen av sykdom (12).

I formildende retning teller aspekter ved pasientens situasjon som forverret forløpet etter helsehjelpsproblemet og reduserte unngåeligheten av dødsfallet, og om det ble iverksatt tiltak for å motvirke skadeomfanget(16). Skjemaet innhenter også vurderinger av hvor mye pasientens livslengde ble forkortet, basert på en skjønnsmessig forventning om levealder for sykdomsbildet ved forsvarlig helsehjelp (12).

Avslutningsvis innhenter PRISM2 vurderinger for variabler som beskriver den enkelte granskerens skåringsprosess, som tidsbruk eller om vurderingene ble hemmet av mangel på subspesialitetskompetanse. Det gjøres også en vurdering av om journaldokumentasjonen var av tilstrekkelig kvalitet til å gjøre vurderingene (Vedlegg 3).

Kun dikotome og kollapsbare kategoriske variabler fra PRISM2- skjemaet inkluderes i analysene for denne oppgaven, med unntak av tidsbruk og administrative variablene *alder og kjønn,* som er høstet fra sykehusregisteret. Instrumentet benytter for øvrig fritekstbeskrivelser, likertskalaer og visuell skala for å innhente granskerens vurderinger (Vedlegg 3)

4.0 Teoridel

I de følgende kapitlene gjøres en sammenfatning av dagens kunnskap innen relevante fagområder for denne oppgaven. Innledningsvis presenteres fenomenene uønskede hendelser og forebyggbare

[2] Acts of commission/ omission fritt oversatt

dødsfall og metoder for identifisering av disse med hovedvekt på journalgjennomgang. Deretter følger et utvalg av temaer fra bedømmelsesteorien og en oppsummering av relevant forskning på helsehjelpsproblemer med journalgjennomgang som identifikasjonsmetode.

2.1 Helsehjelpsproblemer

Spath (2000) og Reynard, Reynolds og Stevenson (2009) bygger på psykologen James T. Reasons (1938-) teorier, og beskriver at helsehjelpsproblemer er resultatet av *uheldige sammenfall av hendelser*[3], der katalysthendelser utenfor en aktørs kontroll kombinerer med såkalte *prestasjonsformende rammefaktorer* og forårsaker uhell(17, 18). De prestasjonsformende rammefaktorene kan være manglende protokoller, personalressurser eller kontrollpraksis som virker desorganiserende på arbeidet[4]. Feil deles gjerne inn i *aktive feil,* som begås av den som utfører en prosess, og *latente* eller *passive feil*[5]*,* som begås av mennesker som ikke deltar i prosessen, eller unngår å utføre en oppgave(17, 19). Forventningen om at menneskelige feil vil forekomme om forholdene tillater det danner grunnlaget for pasientsikkerhetsteorien, og forebyggende strategier i pasientsikkerhetsarbeidet går i stor grad ut på å forenkle og standardisere oppgaver, ettersom det å gjøre feil er bygget inn i menneskets læringsprosess og hvordan vi lagrer informasjon (17).

Definisjonene av uønskede hendelser eller helsehjelpsproblemer innen helsetjenesteforskningen varierer, og en finner også flere definisjoner innen samme publikasjon, som hos Schwendimann et. al (2018) som omtaler både *skade forårsaket av medisinsk behandling*, som tilsynelatende angår kun aktive feil, og *pasientskade assosiert med medisinsk behandling* der også passive feil inngår(20). S*kade forårsaket av helsehjelpen heller enn den underliggende sykdommen, som resulterer i forlenget sykehusopphold, funksjonssvikt ved utskrivelse eller også død* er en mer snever definisjon som Hanskamp- Sebregts et al. (2016) henter fra Werner (2005)*(3)*. Protokollen og skåringsinstruksen for baseres på en relativt vid definisjon, der *ethvert punkt der helsehjelpen falt under en akseptabel standard og førte til skade vil* besrkives som et helsehjelpsproblem (Vedlegg 2, 16).

Amalberti et. al (2009) baserer seg på Batalden et. als (2009) definisjoner, og skiller mellom tre underkategorier av uønskede hendelser(21). De *enkle hendelsene* har overflatiske likheter, og velkjente løsninger som er lette å iverksette. De *kompliserte hendelsene* har felles årsak, men ser ikke nødvendigvis like ut, og har løsninger som må tilpasses konteksten. De *komplekse hendelsene*, har ikke enkeltstående årsaker, dokumentasjonen for disse har gjerne mangler, og flere elementer av samkjørte tiltak må implementeres for å forebygge dem(21).

[3] *"Malevolent coincindence*" fritt oversatt
[4] Mest kjent illustrert ved accident causation model/ «sveitserostmodellen»[4] (Figur 5.1, s. 118 i Reynard, Reynolds, & Stevenson (2009).
[5] Det Spath (2000) omtaler som *active/ passive errors* ser ut til å tilsvare *acts of commission/ acts of omission* hos Hogan (2016)

Amalberti et.al (2009) påpeker at uønskede hendelser som ikke fører til pasientskade, såkalte *near misses* eller *nesten-uhell*, ikke skal inngå i statistikk for uønskede hendelser, ettersom disse kan gi en overberegning av helsehjelpsproblemene(21). Nesten-uhell inngår imidlertid gjerne i klinikerens forståelse av helsehjelpsproblemer, og det ofte oppfordres ifølge Amalberti et al. til å rapportere disse (21). Ehåndboken ved Oslo Universitetssykehus beskriver at «*Hendelser, forhold eller nesten-uhell uten helsemessig betydning (…)»*, skal meldes i sykehusets elektroniske avvikssystem Achilles, av hensyn til økonomiske, tidsmessige eller annet ressurstap(22).

Inklusjonen av nestenuhell i helsehjelpsproblemer er, ifølge Amalberti et. al (2011), et resultat av at en gjerne har sett etter *prosess- relaterte* problemer og *tekniske feil* umiddelbart før hendelsen. Forfatterne foreslår en overgang til et *pasient- relatert fokus*, som utgår fra det utkommet hendelsen får for pasienten, og overordnede taktiske feil i behandlingsforløpet(21).

Amalberti et al.beskriver at de hyppigst forekommende helseskadene som følge av uønskede hendelser er relativt ukompliserte tilstander, som postoperative infeksjoner og liggesår (21). Disse oppstår ifølge Schwendimann et al. (2018) gjerne under rutinebehandling der pasientsikkerheten rammes av kostnadskutt, og er dermed ikke en konsekvens av at helsepersonell underpresterer(20). Slike episodiske hendelser har liten påvirkning på pasientens helhetlige utkomme, men pasienter med kronisk sykdom kan akkumulere helsehjelpsproblemer grunnet avvik i forløpene, som manglende tilgang på spesialister, dårlig koordinering av tjenestene, eller lav etterfølgelse av helseråd(21).

Amalberti et al. mener at de umiddelbare forutgående hendelsene før pasientskaden inntraff vil være et for snevert tidsvindu[6] for deteksjon av problemene, ettersom kompleksiteten i pasientsituasjonen og i behandlingsforløpet fra innleggelsen av påvirker risikoen for helsehjelpsproblemer(21) Dette gjør det mer hensiktsmessig å se på taktiske og organisatoriske valg gjennom hele sykehusoppholdet, et såkalt *utvidet tidsvindu*[7] (17). For journalgjennomgang kan en si at *tidsvinduet* kan begrenses av omfanget og kvaliteten av dokumentasjonen granskerne har tilgjengelig, men at metoden i utgangspunktet omfatter hele sykehusoppholdet, og etter PRISM2 også overordnede strategiske valg gjort i behandlingsforløpet.

[6] In- window timeframe
[7] Out- of window timeframe (19)

2.2 Unngåelige dødsfall

Unngåelige dødsfall er, ifølge Hogan (2016), et enkelt og intuitivt tiltrekkende mål for omfanget av grunnleggende kvalitetsproblemer innen helsetjenestene, som omhandler dødsfall som ikke skulle inntruffet der behandlingen var effektiv og iverksatt i tide (19). Hogan legger i sin fagartikkel *The problem with preventable deaths* fra 2015 frem en rekke argumenter for forsiktighet ved målinger av dette fenomenet som kvalitetsindikator, og forklarer feilkilder som gir variasjon i forkomsten av det (19).

Forfatterens fremste ankepunkt er at tilfeldig variasjon vil ha stor påvirkning på forskjellene i målinger av unngåelige dødsfall for mange spesialiteter, som obstetrikk, psykiatri og elektiv kirurgi, ettersom død et svært uvanlig utkomme for disse(19). I den andre enden av spektret ser en at 40% av dødsfall rammer aldersgruppen over 80 år, og at ¼ av sykehusoppholdene er pasienter over 75 år[8], slik at lokale behandlingstilbud ved livet slutt også påvirker statistikken(19).

Hogan observerer at forskning på unngåelige dødsfall som oftest foregår for et underutvalg som har avgått med døden i helseinstitusjonen eller etter utskrivelse, og mistenker at dette trekker oppmerksomhet fra helsehjelpsproblemer som gir skade hos pasienter som skrives ut i live (19). Videre stilles det i artikkelen spørsmål om hvor robuste målingene av unngåelige dødsfall er, og forfatteren påpeker at selv der helsehjelpsproblemer har inntruffet, er det utfordrende å bestemme hvilken innvirkning disse har hatt ved dødsfall hos eldre, skrøpelige pasienter med alvorlig, kompleks sykdom mot enden av deres naturlige livsløp(19).

Hogan konkluderer sin artikkel med å hevde at unngåeligheten av dødsfall er vanskelig å bestemme, både fordi de som dør i sykehus er et underutvalg som kan antas å ha flere risikofaktorer for helsehjelpsproblemer, og at de aller fleste dødsfall ikke er assosiert med kvalitetsproblemer(19). Forfatteren stiller også spørsmål ved om feilmarginen i det hele tatt kan forbedres ved å justere eksisterende metoder(19). Ifølge Hogan et als studieprotokoll, vil det også være vanskelig å identifisere et enkeltstående helsehjelpsproblem eller hendelsestidspunkt som førte til et forebyggbart dødsfall, da disse mer sannsynlig oppstår som følge av en kombinasjon av problemer i helsehjelpen(16)Dette synet kan sies å gjenspeile James T. reasons *uheldige sammenfall av hendelser* fra forrige kapittel, selv om Reasons teori angår katalysthendelser i foranledningen til en uønsket hendelse.

[8] Nasjonale tall fra NHS / Britiske national health service (17)

Hogan (2016) har den samme bekymringen for påvirkningen av tilliten til helsetjenestene og opplevelsen av risiko ved sykehusinnleggelse hos befolkningen som Amalberti et al uttrykker for helsehjelpsproblemer ved over- eller underestimering forekomsten av unngåelige dødsfalll(19) Ved å benytte slike mål som kvalitetsindikator, vil også de sykehusene som er best til å identifisere egne helsehjelpsproblemer fremstå som om de underpresterer kvalitetsmessig, noe som kan hemme deteksjon og rapportering(19). unngåelige dødsfall kan, ifølge Hogan, ikke beskrive kompleksiteten i helsetjenestene, og avdekker heller ikke de fleste kvalitetsproblemer. Forfatteren konkluderer med at fokuset på unngåelige dødsfall som kvalitetsindikator er ubegrunnet(19).

2.3 Identifikasjonsmetoder

Under presenteres de mest aktuelle identifikasjonsmetodene for helsehjelpsproblemer, med styrker og svakheter ved disse. Avslutningsvis gjennomgås journalgjennomgangsmetoden mer inngående.

2.3.1 Overdødelighet

Ifølge Hogan (2016) var det Florence Nightingale som først så variasjoner i dødsraten på sykehus i London som kvalitetsindikatorer(19). På bakgrunn av helsepolitiske føringer beregnes i dag overdødelighet[9] som standardmål for denne variasjonen ved formelen:

Observerte dødsfall / forventede dødsfall x 100

(19)

Utregningsmetodene baserer seg på sykehusenes administrative data, og skiller seg ved hvordan faktorer som vil gi høyere dødelighet, som alder og komorbiditet, regnes inn i nevneren *foreventede dødsfall*(19). Målenes validitet påvirkes, ifølge Hogan, av variasjoner i kodingspraksis for hoveddiagnose og komorbiditet, og også forskjeller for inklusjonskriterer for dødsfall[10], utskrivningspraksis, og i lokale pallitative tilbud som flytter døende pasienter ut av sykehusene. Som kvalitetsindikator sammenliknes overdødelighet for hvert sykehus med en standardforhold som hentes fra gjennomsnittlig dødelighet i utvalget, noe som forandres fra år til år(19). En politisk drevet tolkning av overdødelighet, er at sykehus i den øvre enden av fordelingen har flere forebyggbare dødsfall(19). Hogan refererer til en studie av Girling, Hofer og Wu i 2012, som viste at så mye som 91% av sykehus som granskes for høy overdødelighet, kan være rammet falske alarmer på grunn av støy i deteksjonsmetoden ettersom forebyggbare dødsfall er sjelden forekommende (19). Verdien av overdødelighet som kvalitetsindikator er omdiskutert, ettersom flere studier viser ingen eller negativ korrelasjon med forekomsten av dødsfall ved

[9] Excess deaths/ excess mortality, andre betegnelser er SMR eller HSMR: standardized mortality rate / hospital standardized mortality rate/SHMI, hospital standardized mortality ration hospital- level mortality indicator
[10] Som dødsfall kun under sykehusoppholdet eller også innen 30 dager etter utskrivelse.

journalgjennomgang(12), Shahian i 2010, også viser at forskjellige overdødelighetsmål produserer radikalt forskjellige rangeringer for det samme utvalget(19) og sammenlikninger, utført av Ifølge Hogan (2016) er det mulig å begrense problemene ved metoden ved å studere overdødelighetsvariasjon innen pasientgrupper der dødsfall er et hyppig utkomme (19).

Overdødelighet er blitt et kjent begrep for allmennhenten under Covid 19- pandemien, som et lettforståelig mål på hvordan dødelighet relatert til COVID- 19 infeksjon og også reduksjon i andre infeksjonssykdommer som følge av smittevernstiltakene, har påvirket nasjonal dødelighetsrate. Ved databasesøk på *Execss deaths*, sees en stor forskningsproduksjon relatert til pandemien. Det kan se ut til at pandemisituasjonen her illustrerer en ytterligere svakhet ved overdødelighet som kvalitetsindikator, ettersom *observerte dødsfall* i enkelte aldersgrupper og for enkelte tilstander kan øke dramatisk som følge av ytre faktorer, også der tilgangen til og kvaliteten på helsehjelpen i det vesentlige er uendret.

2.3.2 Elektroniske avvikssystemer

Etter Spesialisthelsetjenestelovens §3-3 har helseforetakene meldeplikt for hendelser som har ført til eller kunne ha ført til betydelig personskade, og helsepersonellovens § 23 bestemmer helsepersonell kan gi opplysninger videre uten hinder av taushetsplikten når dette er nødvendig for internkontroll og kvalitetssikring (23, 24). Helseforetakene er pålagt å etablere prosedyrer for risikovurdering og avvikshåndtering, og i praksis er det den enkelte ansatte i helseinstitusjonen som melder hendelsen via det interne avvikssystemet(23). Tallene vil være utsatt for målefeil alt ettersom i hvilken grad uønskede hendelser oppdages og rapporteres av helsepersonell under oppholdet, og det bemerkes at det er svært forskjellig praksis for rapportering(19) Sari et al. (2007) demonstrerte i sin journalgjennomgangsstudie at avvik i sykehusets avvikssystem kun detekterte 10% av helsehjelpsproblemene som ble identifisert ved journalgjenngomgang(25), og hos Hibbert et al. (2016) var tilsvarende tall mellom 2 og 8% (26), noe en kan tolke i retning av at avvikssystemene ikke er egnet for beregning av forekomsten av helsehjelpsproblemer eller indikere kvalitetsforskjeller mellom sykehus.

2.3.3 Dødsattester, obduksjon og melding om mulig unaturlig dødsfall.

Gjennomgang av dødsattester eller obduksjonsrapporter brukes ikke systematisk for identifikasjon av forebyggbare dødsfall ved norske sykehus (veileder som referanse?), men ifølge Alfsen et al. (2013) har patologene hatt en uoffisiell rolle som granskere av sykehistorien ved dødsfall(10). Alfsen et al gjennomgikk i 2013 dødsattester og sykehistorier fra kronologiske dødsfall (N= 496) ved Akershus Universitetssykehus, og fant for dette utvalget en underrapportering av mulige unaturlige dødsfall, foruten at det i en femtedel av dødsattestene sannsynligvis ble registrert feil

dødsårsak. Det ble også observert ukorrekt innhold eller ukorrekt innhold og ulogisk oppsett hos henholdsvis 27% og 20 % av attestene. I artikkelen påpekes metodologiske problemer ved dødsattester som identifikasjonsmetode for forebyggbare dødsfall, som at obduksjon kun begjæres der samtykke fra pårørende kan innhentes, og at en ser at stadig færre dødsfall på sykehus fører til obduksjon. Forfatterne konkluderer med at gjennomgang av dødsattester og sykehistorier ved dødsfall på sykehus bør være rutine og håndteres av en definert legegruppe(10).

2.3.4 Journalgjennomgang

Journalgjennomgang er en mye brukt metode for identifikasjon av problemer i helsehjelpen, som av flere forfattere beskrives som en gullstandard for identifikasjon av helsehjelpsproblemer (11, 19) . Journalgjennomgang utført av trenede granskere har, ifølge Norsk Pasientregister, kredibilitet ved at den går inn i kompleksiteten av pasientenes situasjon og helsehjelp (11), selv om det kan oppstå reliabilitetsutfordringer ved både utvalget, granskerne, dokumentasjonsgrunnlaget og måleverktøyet, og metoden er svært ressurskrevende (8, 11).

Planleggingsfasen

I forkant av gjennomgangen, skal det ifølge Håndbok for Journalgjennomgang (2006) holdes et formøte der prosessen beskrives og informasjon som bidrar til at man unngår misforståelser i selve granskningen utveksles (11). Det gjøres ideelt sett et utvalg for en pilotstudie med reliabilitetsanalyse der interrater- reliabilitet er forhåndsspesifisert akseptabel innen en gitt range, som så generaliseres for hele gjennomgangen (27). Alle subjektene skåres av det samme settet granskere i et *fullt krysset design*, eller undergrupper av subjekter skåres av undergrupper av granskere i et *krysset design(27)*. Et fullt krysset design krever at det gjøres flere granskninger, men tillater at systematisk bias mellom kodere kan beregnes og kontrolleres for, noe som reduserer behovet for vanskelig tilgjengelig statistikk (27).

Utvalg

Utvalgsmetodikk og utvalgsstørrelse bestemmes, ifølge Norsk pasientregisters håndbok (2006), av hva resultatene skal benyttes til(11). Om en ønsker å avdekke avvikende praksis innenfor et avgrenset område gjøres journalutvalget på grunnlag av spesialiteter eller hoveddiagnoser der en mistenker avvik, og ukompliserte opphold kan ekskluderes, ettersom det å tilpasse forbedringstiltak har prioritet foran generalisering av funnene(11).

Ifølge Popping (2019) skal utvalggstørrelsen gi en konfidensgrad som tilsier at enigheten er representativ for mønsteret som oppstår dersom hele populasjonen skåres(8). Kraemer (2002) anbefaler et utvalg på 10 til 20% av det fulle utvalget, mens Lacey og Riffe (1996) beskriver at en for dikotome variabler innebærer at en bør ha en utvalgsstørrelse som sikrer enighet for 10

observasjoner innen hver kategori, under antagelsen av at kategoriene har lik preferanse (8, 28, 29). Standardavviket synker raskt etter hvert som utvalgsstørrelsen øker, og flater ut rundt et utvalg på rundt 1000 subjekter(11).Presisjonsnivået blir etter Håndboken for Journalgjennomgang (2006) også kun marginalt bedre om utvalgsstørrelsen økes utover 1500, mens kostnaden øker med utvalgsstørrelsen, allikevel skal utvalget være representativt og tidskonsistent for den institusjonen som undersøkes(11).

Granskerutvalg

Dersom både subjekter og granskerne er tilfeldige utvalg fra større populasjoner, er det, ifølge Hallgren (2012) mulig å generalisere resultatene[11] (27). Der subjektene er tilfeldig utvalgt, men granskerne er fastsatte, kan ikke resultatene generaliseres, noe som er avgjørende for tolkningen, men ikke reliabilitetsberegningen(27)

Ifølge Håndboken for journalgjennomgang, bør minimum to granskere med nødvendig kompetanse gjennomgå samtlige journaler(11)[12], mens Mchugh (2012) bemerker at en trenger minst 5 granskere for å oppnå 90 % reliabilitet[13] noe som for de fleste tilfeller vil være en urimelig ressursbruk(30). For å kunne generalisere til den større befolkningen av granskere tilgjengelig, skal granskerne, ifølge Popping (2019) være utskiftbare. Dette innebærer at det gjøres et tilfeldig utvalg av granskere fra den større granskerpopulasjonen som antas å være like i evner og erfaring, samt at disse gjennomgår den samme kursingen i bruk av måleverktøyet og utfører skåringene uavhengig av hverandre[14](8). Ved å implementere slik kursing reduseres ifølge Mchugh (2012) variabiliteten i hvordan dokumentasjonsgrunnlaget tolkes og dokumenteres(30)

En beskrivelse av hvilken kursing granskerne gjennomgår er svært viktig ved komplekse skåringsoppgaver, men som regel er ikke detaljene for denne kursingen tilgjengelig(8). For å sikre legitimitet skal granskerne ha god kjennskap til kodingsstandardene, og ikke være involvert i de kliniske vurderingene som granskes(11). Popping beskriver at en kan bruke en ytterligere gransker eller gruppe granskere for å vurdere journaler der det er uenighet mellom granskerne(8). Hos Hogan et al (2015) brukes slike ledende granskere for ekstra gjennomgang også av alle journaler der det er identifisert helsehjelpsproblem, uavhengig av enighet mellom granskerne (12).

[11]Med *Random effects model (22).* Modellen angår ikke våre forskningsspørsmål og omtales ikke.
[12] Her er granskerne det Popping Kaller *Random effects* (Popping 2019), der de ikke er tilfeldig utvalgt er granskerne *Fixed effects*.
[13] Påstanden om 5 granskere for 90% begrunnes ikke av Mchugh (2012).
[14] Dvs at en måler intrarater- reliabilitet for ekspertgranskeren istedenfor inter- rater reliabilitet mellom granskerne, disse begrepene omtales senere.

Dokumentasjon

Helsepersonellovens § 39 pålegger den som yter helsehjelp å nedtegne eller registrere relevante og nødvendige opplysninger i en pasientjournal(23), som har til hensikt å bidra til pasientsikkerhet og etterprøvbarhet(11).Etter dokumentasjonsplikten skal journalen foruten administrative, sosiale og demografiske data, inneholde pasientens diagnoser, observasjoner og funn, behandling og annen oppfølging, som sakkyndige uttalelser, innelggelsesbegjæringer og epikriser(11).

En viktig del av granskningen vil, ifølge Håndboken for Journalgjennomgang, være vurderingen av dokumentasjonskvaliteten, og det anbefales at granskerne beskriver disse manglene ved forklaringsfaktorer som inkonsistente eller uklare formuleringer, avvik mellom administrative data og journal, manglende hoved- og bidiagnosekoder, eller at dokumentasjon i journalen har blitt oversett(11). Videre bemerkes at en nærhet til journaldokumentasjonen og helsepersonell som har bidratt til denne, er en fordel dersom det oppstår uklarheter eller spørsmål (8). Det vil ligge nærliggende å tenke at en slik nærhet også forutsetter en uavhengighetsvurdering, om granskerne kommer fra det samme kollegiale miljøet og også om det som måles er kvaliteten på klinikernes helsehjelp.

Skjemaverktøyet

Måleverktøy som benyttes bør, ifølge Hallgren et al., vurderes for faktorer ved målenivå av variablene og definisjonsrom for disse som kan påvirke reliabilitetsmålene(27). Skalaer som har prestert dårlig i foregående studier, vil antagelig fortsette å produsere lav IRR, og en begrensing av range[15] når nye populasjoner skåres kan også gi lav IRR, også for skalaer som har prestert godt tidligere(27). Pilottesting er egnet for å vurdere om nye eller modifiserte skalaer er passende, og dersom variabler utelukkes må dette begrunnes(27).

Ved bruk av skjemaverktøy ved journalgjennomgang følges *a posteriori* koding, som etter Montgomery og Crittenden (1977) i Popping (2019) innebærer at kategoriene settes før skåringsprosessen begynner[16], og granskerne tilskriver enhetene til kategoriene uten frihet til å forandre disse(8). Ifølge Popping må kategorier må være uttømmende og gjensidig utelukkende beskrivelser av det samme karakteristika, og dersom de ikke er uttømmende, bør dette ifølge Popping løses ved å legge til en "annet" kategori slik at en unngår at det skåres i den nærmeste kategorien. Der variabelen er kompleks, bør forskeren gi beskrivelser av hvilke faktiske virkelighetspåstander som er tilstede(8).

[15] For kategoriske variabler betyr dette at enkelte kategorier ikke benyttes (14)
[16] Motsatt a Priori koding, som ved pilotstudier der målet er å utvikle et skjemaverktøy ifølge Montgomery og Crittenden 1978 i Popping (2019).

Ifølge Bennett (1954) blir kodingen mer kompleks når skjemaverktøy med konstruktbeskrivelser brukes(31). Konstruktene tilhører ikke subjektenes natur, men er kvaliteter som utvikles av forskere eller brukere av verktøyen granskerens subjektive tolkninger av disse gjør seg gjeldende til tross for enighet om kodereglene(31) Popping (2019) tilfører at det også bør medfølge en vurdering av forskjeller i de uformelle subjektive skåringsprosedyrene hver gransker benytter, men at dette i praksis ikke gjøres (8).

2.4 Bedømmelsesteori og bekreftelsesfeil for ekspergranskeren

Mye av det eksperimentelle arbeidet som danner grunnlaget for *bedømmelsesteorien* eller *heuristikken*, ble utført av forskerne Tversky og Kahnemann på 1960 og 1970- tallet, som på bakgrunn av eksperimentene utviklet teorier om ubevisste kognitive prosesser som reduserer kompleksiteten i bedømmelsesoperasjoner(32). Disse prosessene gjør det enklere å konkludere om sannsynligheten for en usikker hendelse, men kan også føre til alvorlige og systematiske feilslutninger (32).

Tversky og Kahnemann presenterer heuristiske i sin artikkel *Judgment under Uncertainty: Heuristics and Biases* fra 1974 prinsippene *Representativhet og Feiloppfatninger av tilfeldighet*, og *Gamblerens bekreftelsesfelle(32).* Artikkelen beskriver at mange sannsynlighetsvurderinger er beslutninger om hva sannsynligheten er for at *hendelse A* stammer fra *prosess b*. *Representativhet* innebærer at granskeren konkluderer at dersom A sees som svært representativ for B, er sannsynligheten for at A har sin opprinnelse i B høy, og motsatt dersom A ikke sees som representativ for B(32). Dette kan være misvisende, ettersom det at fenomener har en liknende framtoning ikke nødvendigvis indikerer en sammenheng mellom disse(32). Med representativhetsprinsippet følger også en v*aliditetsillusjon*, eller en tiltro til egne vurderinger som avhenger av graden av representativheten(32).

Tversky og Kahnemann demonstrerte representativhetsprinisppet ved at forsøkspersoner som fikk karakterbeskrivelser av et subjekt som ble assosiert med en gitt yrkesgruppe, tendenserte mot å konkludere med at vedkommende hadde et slikt yrke, uavhengig av kjennskapen til at forekomsten av yrkesgruppen i befolkningen var svært lav[17](32). Der subjektene ble beskrevet nøytralt og skulle plasseres i ett av to yrker, ble også halvparten av subjektene plassert i hver yrkesgruppe, til tross for den skjeve forekomsten. Forfatterne konkluderte med at beskrivelser av representativhet for en egenskap, og også unyttig informasjon, påvirker konklusjonene uavhengig av kjennskap til fordelingen(32).

[17] *"Steve is a librarian"*- forsøket (25)

Tversky og Kahnemann demonstrerte også intuitive *feiloppfatninger av tilfeldighet* hos forsøkspersonene, som at Kron/ Mynt - sekvensen K-M-K-M-M-K ble antatt å være mer sannsynlig enn M-M-M-K- K- K, ettersom den siste ikke framstår som like tilfeldig(32). Sekvenser der kron og mynt ikke har like mange hendelser ble også vurdert å ha mindre sannsynlighet for å være tilfeldig, uavhengig av at det var et lite antall kast. Forfatterne forklarer dette med at forekomsten ikke representerte «rettferdigheten» i et myntkast, og omtalte det som *Gamblerens bekreftelsesfelle*[18]. Prinisppet er beslektet med både representativhet og manglende hensyn til utvalgsstørrelse, der lite utvalg som har liknende forekomst av et trekk som befolkningen antas å være trukket fra denne, mens dette ikke antas for et større utvalg med samme tendens, men mindre likhet. Dror (2020) beskriver også at eksperten kan ha en forventning om forekomst[19] basert på foregående erfaring, og at det vil være vanskelig å gjenkjenne et karakteristika hos nye personer om det har vært sjeldent forekommende hos personer vedkommende har erfaring med(33).

En annen kjent effekt som kan gjøre seg gjeldende i forskning er *Hawthorne- effekten,* som ifølge Sedgwick og Greenwood (2015) er et velkjent fenomen som opprinnelig utgår fra forsøk der en påviste at forsøkspersoners atferd tilpasset sin atferd til forsøkets hensikt om de visste at de ble observert(34)..

Dror (2020) sammenfatter også feilslutninger som utgår fra kognitive bekreftelsesfeil, og som i særlig grad rammer ekspertgranskeren(33). Forfatteren innleder med beskrivelser av hvordan en gjerne mener at om kognitive bekreftelsesfeil rammer ekspertgranskeren, er det mer sannsynlig at dette forklares som moralske eller kompetansemessige mangler, ettersom det antas at ekspertgranskeren er immun mot bekreftelsesfeil. Slike forutantagelser avvises av Dror, som argumenterer for at ekspertgranskeren har større risiko for enkelte typer bekreftelsesfeil, ettersom deres erfaring og kunnskap lar dem bruke selektiv oppmerksomhet og sammenfatting av informasjon, og innta en *ovenfra og ned- tilnærming* der forventninger og antagelser kan oppstå(33). Reynard, Reynolds og Stevensons (2009) beskrivelser samtidig at eksperter gjerne benytter en *ferdighets - og regelbasert problemløsning* og har høyere risiko for å begår «slips and lapses», fordi sekvensene gjennomføres på en ubevisst måte(18).

Dror (2020) erfarer at ekspertens bekreftelsesfeil kan fremprovoseres av *datagrunnlaget* eller *referansemateriell,* der granskeren ser bort i fra enkelte funn ettersom det for øvrig ikke passer til antagelsene hun har gjort på bakgrunn av den øvrige informasjonen(33). Dette kan også påvirkes av *kontekstuell informasjon*, som at kunnskapen om et subjekts bakgrunn kombineres med tidligere erfaring med mennesker med den samme bakgrunnen i liknende situasjoner og forårsaker

[18] Gamblers fallacy
[19] *Base rate bias*

feilslutninger (33). Disse feilkildene kan minne om Tversky og Kahnemans *representativhet,* ettersom manglende likhet kan gjøre at en overser eller feiltolker informasjon(32). Dror (2020) kaller en *snøball-kaskadeeffekt*, som kan oppstå der granskerne samarbeider,

Bekreftelsesfeil relatert til *lojalitet og minside-forutsetninger*[20] eksemplifiseres av Dror ved at en ser at rettsmedisinske eksperter gir uttalelser til fordel for den siden som har hyret dem, uavhengig av de rettsmedisinske bevisene(33). Hierarkiske systemer eller organnisasjonsmessige pressfaktorer, kan ifølge forfatteren påvirke personer i det samme arbeidsmiljøet til å levere det produktet som forventes, og utanning og erfaring kan påvirke hvem en opplever et lojalitetsforhold (33).

2.5 Forskning på helsehjelpsproblemer med reliabilitetsberegning

Under presenteres forskning med journalgjennomgang som identifikasjonsmetode for helsehjelpsproblemer, med fokus på studier der en beregning av interrater- reliabilitet inngår. Det ble ikke funnet sammenliknbare studier der Gwets AC1 var beregnet, og dette er en svakhet ved gjennomgangen.

Hayward og Hofer (2001) undersøkte forebyggbarheten av dødsfall i 2001 for et utvalg (n=111) fra 4198 dødsfall på forskjellige Veteran Affairs (VA) sykehus(35). Forebyggbarheten ble vurdert på en 5 punkts ordinal skala samt ved prosentanslag av sannsynlighet for overlevelse 3 måneder etter utskrivelse med optimal helsehjelp. Reliabilitet for forebyggbarheten vurdert ved Intraclass Coefficient var mellom .24 og .34 (35). Utvalget ble gjort stratifisert tilfeldig med overutvalg av grupper med økt risiko for forebyggbart dødsfall (n=101, 56%). Terminalt syke (n =66) ble ekskludert fra utvalget. Snittalder for død i utvalget var 69år SD (35). Sannsynlig forebyggbart dødsfall ble funnet i 6% av journalene, mens mulig forebyggbarhet ble funnet i 22,7% av journalene(35). Forfatterne bemerker at de samme granskerne som vurderte et mulig forebyggbart dødsfall, også anslo sannsynligheten for overlevelse med forsvarlig helsehjelp som gjennomsnittlig 20% for disse pasientene, og 43 % for pasienter der dødsfallet ble vurdert sannsynlig forebyggbart. Ekstremskårende granskere ble funnet å påvirke dataene ved gjennomsnitt/ typetall sammenlikning, samt at de fleste positive funn representerte outliere, der de fleste granskerne vurderte sannsynligheten for forebyggbart dødsfall som lav(35)

Reliabilitet ble i Hayward og Hofers studie beregnet for et lite utvalg (n=35) av den totale populasjonen (N = 4198)(35), og usikkerhetsestimat er ikke oppgitt, noe som påvirker overføringsverdien av en allerede lav IRR. For de ordinale variablene er det ikke beregnet IRR, noe en kan se for seg at kan begrunnes ved at variablene er det samme konseptet målt på to forskjellige måter, men dette er ikke forklart. Kursing og blinding er godt beskrevet, og det er angitt at

[20] *Allegiance og myside- biases*

gjennomsnittlige skår ikke varierte signifikant i granskerutvalget eller gjennom studieperioden(35). Forfatterne utførte en Monte Carlo- simulering, og fant at forutsatt minst 100 granskere, ville minst èn gransker vil finne forebyggbarhet (35). Det savnes en vurdering av om antagelsene for simulering er oppfylt, og også hvordan en unngår «søppel inn, søppel ut», ettersom en lav IRR innebærer stor variasjon i dataene. Utvalgsmetodikken, der det er gjort et overutvalg av enkelte journaler, virker relevant etter begrunnelsen om økt forekomst i denne undergruppen, men det undersøkes ikke om journalene det er gjort et overutvalg for også har høyere forekomst i denne studien[21].

Hanskamp- Sebregts et al. (2016) inkluderte i sin forskningsoppsummering av 24 journalgjennomgangstudier med reliabilitetsvurdering, og fant at forekomst av uønskede hendelser var 2.9%til 18% og unngåelige dødsfall var henholdsvis 1% til 8.6% der metoden fulgte strukturert journalgjennomgang med implisitt bidrag fra Brennan et als (1991) Harvard Medical Practice Study (36), som er mest relevant for vår studie. Gjennsnittlig IRR for det kombinerte artikkelutvalget var henholdsvis 0,65 og .055 for disse to metodene(3). IRR var signifikant høyere når studiene ikke inkluderte mer enn fem granskere(3).

Forfatterne oppgir en antagelse om at IRR vil være høyere for studier med et lite antall granskere, uten at dette begrunnes. Antagelsen samsvarer også med funnene i studien(3), men motstrider prinsippet om *regresjon mot gjennomsnittet*, det vil si at mindre utvalg vil ha større variasjon enn et større, og også Mchughs (2012) påstand om at en må ha 5 granskere for å få en reliabilitet over 90%(30). Kanskje kan påstanden begrunnes av at en ved få granskere har bedre mulighet til å sikre like egenskaper for disse,og dette vil gjøre det ufordrende å gjøre et tilfeldig granskerutvalg, det vil si at prinisippet om regresjon mot gjennomsnittet ikke gjelder. usikkerhetsestimat for IRR er heller ikke oppgitt i denne studien. Forfatterne utgår selv fra èn oppgitt definisjon av uønskede hendelser for studien, men beskriver ikke om det er vurdert om de inkluderte studiene følger samme definisjon, noe som kan bety at studiene måler noe forskjellige fenomener.

Lilford et al. (2007) gjennomførte en utvidelse av Goldmans litteratursammenfatning av journalgjennomgang fra perioden 1959-1991, med ny inndeling etter om metoden var implisitt eller eksplisitt[22]og om det var årssakssammenhenger, ved aktive og passive feil, eller utkomme i form av uønsket hendelse som ble målt(37). 26 artikler ble inkludert for samvarasjonsanalyse ved ANOVA[23] mellom de oppgitte IRR- målene og fordeling, metode og utkomme. Forfatterne bekreftet sin hypotese om at kappa ligger høyere der utkomme måles istedenfor årsakssammenheng ($p < .008$). Laveste kappa var .32 for implisitte gjennomganger der årsakssammeng ble målt, og

[21] Forutsatt at det ikke er skrevet andre artikler for den samme studien.

[23] Analysis of Variance

høyeste var .70 for eksplisitt gjennomgang for deteksjon av utkomme, og forfatterne foreslo på bakgrunn av funnene at journalgjennomgang bør inkludere eksplisitte spørsmål kombinert med et avsluttende implisitte vurderinger.

Lilford et al. påpeker at kappaparadoksene kan inntre ved datafordelingen i studien, og bemerker at en svakhet ved studien er at koeffisientvalget var begrenset, ettersom Cohens kappa var mest brukt i artikkelutvalget (37).Forekomsten av helsehjelpsproblemer for de inkluderte studiene er ikke presentert i artikkelen, men er vedlagt i tabell. Forfatterne utgår fra beregnet kappa i artiklene, sorterer disse etter metode og utkomme, og beregner mean Kappa innen gruppene. Antall journaler per IRR-beregning er oppgitt, men ettersom mean beregnes for oppgitt Kappa per studie, vil en Kappaberegning for få journaler ha like stor innvirkning på analysen som større gjennomganger. Mange av artiklene som er inkludert er svært gamle, og en ser også at det inkluderes to- stegs journalgjennomganger der kappa beregnes i begge omganger. Her kan en forvente at kappa i 2. steg, der leger gjennomgår journaler som allerede er underutvalg av de sykepleiere gjennomgikk, vil ha høyere reliabilitet[24]. Inklusjonen med en såpass annerledes metodikk er ikke begrunnet, og kan tyde på at det også kan være endel metodevariasjon og utvalgsvariasjon i artikkelutvalget. Dette kan være problematisk også fordi det ikke benyttes rådata i beregningen av IRR, men beregnes gjennomsnittlig IRR for IRR- funn fra artikkelutvalget. Som Hallgren (2012) beskriver vil også gjennnomsnittet av flere målinger være mer reliable enn enkeltmålinger (27), noe som gjør at en på denne måten antagelig får høyere IRR enn om en hadde beregnet denne for rådataene samlet. Dette prinsippet henger også sammen med prinsippet om regresjon mot gjennomsnittet, som ble nevnt for Hanskamp- Sebregts et als (2016) artikkel når det gjelder vurdering av et størrelsen på granskerutvalget.

Brennan et al. (1991a, 2004b) gjennomførte på 1980- tallet den mye omtalte Harvard Medical Practice Study, ved 51 somatiske sykehus i New York. Studien var del av en interdisiplinær studie og formålet var å utvikle mer oppdaterte metoder og reliable estimater av forekomsten av uønskede hendelser og uaktsomhet, og metodikken utviklet i studien adopteres av forskere ved journalgjennomgang også i dag. I første omgang ble et tilfeldig utvalg (n=30 195, N=2 671 863) gjennomgått av sykepleiere og journalanalytikere[25] med tanke på uønskede hendelser. De markerte journalene (n=7817) ble gjennomgått av uavhengige legegranskere, og vurdert for sannsynligheten for uønsket hendelse på en 0- 6 skala der >1 ble vurdert som postivt skår, samt en tilsvarende skala der skår 4 eller mer ble vurdert som uaktsomhet. og det ble påvist uønsket hendelse i 3.7 % og uaktsomhet i 1 % av journalene. 70.5 % av hendelsene ga kortvarig funksjonsnedsettelse, 2-6 %

[24] Wilson 1995 (i 30)
[25] Det fremgår ikke tydeligere hvilken yrkesgruppe dette er.

forårsaket permanente skader og 13.6 % død. Forekomsten av uønskede hendelser økte markant med økende alder, og uaktsomhet var mer vanlig hos pasienter som var utsatt for mer alvorlige hendelser. for kjønn var forskjellene ikke signifikante. Funnene ble ved ekstrapolering med statistiske vekter overført til sykehusbefolkningen og det ble beregnet at potensielt kunne det kan ha inntruffet 27 179 uaktsomme hendelser, som kan ha ført til 2550 permanente funksjonsnedsettelser og 13 451 dødsfall i studieperioden. Harvard- metoden som ble utviklet for Brennan et als studie er, ifølge Hibbert et al (2017) mest brukte tilnærmingene til beregning av helsehjelpsproblemer i tillegg til Global Trigger Tool(26)

Validiteten ble i Brennan et als studie ble testet ved at en medisinsk journalanalytiker opererte som gullstandard og revurderte 1 % av journalene. Beskrivelsen av validitetstesten likner kanskje mer et reliabilitetseksperiment, ettersom det kun er yrket til granskeren som byttes ut, og metodene ellers ser ut til å være like. Reliabilitet ble undersøkt ved en gjentatt vurdering av et team bestående av legespesialister og journalanalytikere. Her kan en argumentere for at dette ikke er egnet som reliabilitetseksperiment, ettersom granskeren og granskerteamet ikke er sammenliknbare i evner og erfaring, og skåringsprosessen ikke er lik. En savner en vurdering av hvordan gruppedynamikken kan ha virket inn på reliabiliteten. Cohens Kappa var .610 for uønskede hendelser og .240 for uaktsomhet, og forfatterne tolket dette som at legene var ofte uenige om vurderinger av om helsehjelpsstandarder var møtt for de journalene de var enige om at det hadde inntruffet uønskede hendelser. Det er ikke oppgitt usikkerhetsestimat for kappa, men ettersom utvalget er stort, vil en antagelig kunne gå ut ifra at beregningen er signifikant og at nedre KI ligger nær beregnet Kappa.

Forfatterne eksemplifiserer flere ikke-uaktsomme hendelser ved at pasienter kan ha en uønsket reaksjon på legemidler, selv om disse er forsvarlig forordnet. Slike eksempler kan synes å være hendelser som ikke reelt kan forebygges, og det overveiende flertallet av uønskede hendelser identifisert i Brennan et als studie er ikke- uaktsomme. En leter i Brennan et als artikler etter en begrunnelse for behovet å identifisere slike hendelser, med tanke på hvor ressurskrevende journalgjennomgangsmetoden er, dersom hendelsene ikke forårsakes av kvalitetssvikt i helsetjenestene.

Forfatterne fremhever at uønskede hendelser ikke nødvendigvis indikerer dårlig helsehelpskvalitet, og påpeker at medisinske journaler dessuten er en dårlig kilde for å detektere uaktsom praksis som ikke gir skade. Forfatterne bemerker at mange pasienter hadde svært alvorlig underliggende sykdom, og at legegranskerne ikke vurderte hvor mye livslengden ble forkortet der uønsket dødsfall forekom, men det er noe usikkert om forfatterne begrunner å ha utelatt en slik vurdering, eller mener det er en begrensning ved egen studie.

PRISM studiene For å undersøke korrelasjon mellom overdødelighet[26] og forebyggbare dødsfall identifisert ved journalgjennomgang, ble det av Hogan et al. i 2009 og 2014 -2015 utført studier med navn Preventable Incidents, Survival and Mortality Studies ved 25 engelske sykehus, basert på metodikk fra blant annet Hayward og Hofers (2001) tidligere nevnte studie.(12). Obstetriske, psykiatriske og pediatriske pasienter ble utelatt fra utvalget, før det ble gjort et tilfeldig utvalg journaler fra dødsfall på sykehus i den nasjonale Engelske helsetjenesten NHS (n=1000). Journalgjennomganger ble utført av legegranskere og 10% av journalene ble undersøkt for interrater-reliabilitet. Et skjema for journalgjennomgang ble utviklet under PRISM- studiene, og videreutviklet til skjemaet PRISM2 som benyttes i vår studie (vedlegg 3). Hogan et al. Konkluderte med at 13.1 % av journalene var utsatt for helsehjelpsproblemer, og at forekomsten av dødsfall som hadde 50 % eller større sjanse for å være forebyggbare var 5,2% og interrater- reliabilitet var innen kvalitativ tolkning *moderat* med Kappa henholdsvis 0.54 og 0.49(12).

Den norske oversettelsen av skjemaet utviklet i Hogan et als gjennomganger benyttes i vår studie, og Hogan et als resultater er svært interessante som sammenlikningsgrunnlag for vår studie. Aspekter som allikevel kan påvirke forskjeller i utkomme for helsehjelpsproblemene i disse to studiene vil være eksklusjon av psykiatriske pasienter og bruk av en lead reviewer, som en ser for seg kan øke beregningene av reliabilitet. Det er ikke beskrevet om det ble undersøkt en slik effekt for journalene som gjennomgikk lead review i Hogans Studie.

Deilkås et al. (2015) oppsummerte resultater fra det nasjonale pasientsikkerhetsprogrammet fra perioden 2010 to 2013, der 18 offentlige og 5 private offentlige sykehus rapportere forekomsten av helsehjelpsproblemer basert på journalgjennomgang med Global Trigger Tool (GTT)(38). Opphold med kortere varighet enn 24 timer samt opphold fra pediatriske, psykiatriske og rehabiliteringsopphold ble ekskludert og et tilfeldig utvalg (n=40851) fra 2 249 957 ble gjennomgått av et lag mediske eksperter. Dataene ble behandlet som tidsserier etter alvorlighetstype og endringer i av helsehjelpsproblemer per ble beregnet. Beregnet forekomst av de alvorligste hendelsene ble signifikant redusert fra 16.1% i 2011 til 13.0% i 2013, reliabilitet ble ikke beregnet for metoden(38).

Deilkås utførte også i 2017 i et samarbeid med svenske forskere en sammenlikning av tallene fra Deilkås et al (2015) og tilsvarende tall med Global Trigger tool fra svenske sykehus, og fant at det ikke var signifikante forskjeller i forekomsten av helsehjelpsproblemer mellom de to landene. Enkelte forskjeller ble imidlertid identifisert i hvilke type hendelser som ble hyppigst identifisert og hvilke pasientgrupper som ble rammet av disse(39)

Deilkås et als studie er svært interessant med tanke på tendenser for helsehjelpsproblemer ved norske sykehus,ettersom utvalget er stort og sammenfatningen av dataene er gjort på en oversiktig og etteretterlig måte. Effekten av at målingene er gjort med journalgjennomganger som er lovpålagte og forskjeller i skåringsprosessene mellom forskjellige sykehus er ikke omtalt utover at det samme instrumentet ble brukt, og en ser at Hibbert et al. (2016) som inkluderte 44 studier fra 16 land for å gjennomgå anvendelsen av Global Trigger Tool, konkluderte med at ettersom forekomsten av helsehjelpsproblemer identifisert med dette instrumentet varierte mellom 7 og 40% i studiene, er resultater fra anvendelsen av GTT ikke egnet til sammenlikning mellom institusjoner.

Rogne et al (2019) utførte journalgjennomganger på 1000 kronologiske ikke- psykiatriske dødsfall hos pasienter over 16 år et St. Olavs hospital(5). Median alder i utvalget var 77år og kvinner utgjorde 45%: Legeekspert stod for gjennomgangene og kursing ble gjennomført som flere vekselsvis individuelle gjennomganger av mindre journalutvalg som så igjen ble gjennomgått i plenum for å sikre likhet i måte granskerne tilnærmet seg skåringene. Det ble benyttet mortalitetsindeksskåringer[27], sammenlikninger av innleggelsesdiagnose og dødsårsak, identifikasjon av helsehjelpsproblemer ved enhver hendelse der helsehjelpen ikke møtte faglige standarder, og indikasjoner på overbehandling for å konkludere med helsehjelpsproblem. Forfatterne konkluderte med at 4.2% av dødsfallene hadde mer enn 50% sannsynlighet for å være unngåelige, og at disse pasientene fikk livet forkortet med minst 1 år. Svært få av dødsfallene var rapportert I sykehusets avvikssystem, og forfatterne anbefaler jevnlige journalgjennomganger av kronologiske dødsfall uten å gjøre utvalg som en forbedring av de mest brukte metodene(5). Kursing er godt beskrevet og virker hensiktsmessig i Rogne et als studie. Interrater-reliabilitetsvurdering kunne vært svært interessant for å vurdere effekten av kursingen, og også fordi forfatterne anbefaler et nytt syn på utvalgsmetodikken for fremtidige studier, men IRR er dessverre ikke beregnet.

Lipczak og Schiøler (2001) gjorde proporsjonale utvalg av 1092 sykehusinnleggelser 17 sykehus gjennom det Danske Pasientregisteret og utførte journalgjennomganger ved en tre- trinns prosess, der sykepleiere screenet journalene etter 18 risikokriterier før leger utførte uavhengige journalgjennomganger(40). Ved uenighet i det andre trinnet gjennomgigg ytterligere to leger journalene og som til slutt konfererte med hverandre. Forfatterne fant helsehjelpsproblemer i 9 % av innleggelsene, og at 40% av hendelsene var unngåelige og at de fleste rammet aldersgruppen 76- 96 år. Permanent funksjonsnedsettelse eller død ble identifisert i 2.9% av journalene(40).Ettersom uenighet ser ut til å være utelukket ved metoden har det ikke vært hensiktsmessig å beregne IRR i

[27] Charlosn mortality index

denne studien. Det skilles ikke mellom permanent funksjonsnedsettelse og død, og kriteriene for funksjonsnedsettelse er ikke beskrevet.

3.0 Måleteori

I dette kapitlet presenteres definisjoner av validitet og reliabilitet, og forholdet mellom interrater-reliabilitet, og tilfeldig enighet beskrives.

3.1 Validitet

Ifølge klassisk test teori består enhver måling av en sann verdi og en feilkilde som gjør seg gjeldende som en diskrepans fra verdien målt ved en referanseskår eller gullstandard, det vil si en målemetode som antas produsere sane verdier (8, 41). En målings validitet angår, ifølge Popping (2019) og Gwet (2021), hvorvidt funn i en undersøkelse representerer sanne verdier, og for å vurdere validitet må også en referanseskår være tilgjengelig (8). Validitet måles, ifølge Gwet (2021), med validitetskoeffisienter, men om det ikke er en referanseskår tilgjengelig verdifull informasjon om validiteten (41). For reliabilitetsstudien i vår undersøkelse er ingen referanseskår tilgjengelig, men det flere validitetsaspekter som må gis oppmerksomhet. *Utvalgsvaliditet* er, ifølge Popping, hvorvidt tilgjengelig data er et tilfeldig og representativt utvalg for befolkningen som skal undersøkes, slik at dataene kan sees som statistisk representative(8), og Laake et al (2007) beskriver en manglende representativitet av utvalget som *utvalgsskjevhet*. Denne utgjør sammen med *informasjonsskjevhet*, som omhandler feil i datagrunnlaget, og *statistisk validitet*, som omhandler rett bruk av effektmål og statistiske tester, som utgjør den *interne validiteten(42)*.

Ekstern validitet er videre knyttet til resultatenes *generaliserbarhet*, som ifølge Laake et al i hovedsak bestemmes av design samt tilfeldige og systematiske feilkilder(42). I hvilken grad et skjemaverktøy produserer sanne skår, omtales av Popping som *konstruktvaliditet(8).* For variabelen som måles vil beskrivelser for kategoriene reflektere deres meninginnhold, noe som reduserer tekstens tvetydighet og sneverer inn tolkningsrommet snevres inn, noe Krippendorff (1980) omtalte som *semantisk validitet* (8, 43). Popping og Krippendorffs validitetsdefinisjoner tar utgangspunkt i innholds- og tekstanalyse og dermed kvalitative forskningstradisjoner, mens Laake et al. utgår fra den epidemiologiske forskningen.

3.2 Validitet som spesifisitet og sensitivitet

Hayward et al. (2007) beskriver i sin metodeartikkel et velkjent problem innen epidemiologien, der målemetoder med moderat reliabilitet vil gi en vesentlig overberegning ved inferens av statistikk fra

et utvalg til befolkningen dersom utkommet i utgangspunktet har lav forekomst(44). Fenomenet eksemplifiseres ved at en metode med 90% sensitivitet og 90% spesifisitet, vil identifisere 45 korrekte positive og 995 falske positive for et utkomme med en forekomst på 0,5% i en befolkning på 10 000, og dermed gi en dramatisk overestimering av utkommet.

Artikkelen argumenterer for at ettersom vi ikke kjenner sensitivitet eller spesifisitet for journalgjennomgangsmetoden, og metoden også er kjent for å gi lav til moderat reliabilitet, bør en ha mindre tillit til gjennomsnittlige skår som vil gi feilklassifiseringer og overvurderinger av variansen mellom subjektene der en har få granskere(44). Forfatterne beskriver at det å vurdere prevalens for en befolkning på bakgrunn av et utvalg, krever en metode med nær perfekt spesifisitet, noe som innebærer nær perfekt reliabilitet for metoden sammenliknet med en gullstandard(44).

Videre kritiseres tilnærmingen der en enighet mellom majoriteten av granskerne for kategoriske variabler bestemmer forekomsten av forebyggbare dødsfall, og mener at denne heller bør erstattes med prosentvurderinger at sannsynlighet for overlevelse med optimal helsehjelp(44). Forfatternes sistnevnte argument begrunnes ikke utover at unngåelighet "mer naturlig" ligger på en 0- 100 skala, til tross for at standardene helsehjelp vanligvis måles mot i lovverket og opp mot faglige standarder utgår fra hvorvidt helsehjelpen kan kategoriseres som forsvarlig og en kan argumentere for at dette betyr at standardene ikke ligger på en kontinuerlig skala. Heuristikken og James T. Reasons teorier, som disse er beskrevet tidligere, antyder at menneskelig hukommelse og bedømmelse også opererer mer «kategorisk» enn kontinuerlig, og en kan savne en bedre begrunnelse for Hayward et. als påstand.

3.3 Reliabilitet, enighet og uenighet

Reliabilitet sees av Popping (2019) som en bestanddel av *semantisk validitet* og er en nødvendig, men ikke tilstrekkelig, forutsetning for validitet(8). Reliabilitet setter en grense på validiteten av en måling, og uten perfekt reliabilitet kan ikke målingen bli valid(8). Laake (2007) beskriver reliabilitet som i hvilken grad gjentatte målinger gir det samme resultatet(42).Denne definisjonen er svært lik poppings definisjon av *reproduserbarhet(8)*, mens Tinsley og weiss (2000) forholder seg til reliabilitet som en indikasjon på i hvorvidt *variasjon* i skåringer kan tilskrives forskjeller mellom subjektene som granskes(45). Laakes reliabilitetsdefinisjon er til forveksling lik Tinsley og Weiss` enighetsdefinisjon, der *enighet* er granskeres tendens til å tilskrive like skår for de samme subjektene, men Laakes definisjon angår målinger der Tinsley og Weiss omtaler skåringer (42, 45).

Temaene *stabilitet, reproduserbarhet og riktighet*[28] inngår ifølge Popping i reliabilitetsbegrepet. Bauer et al (2000) beskriver at stabiliteten sees ved reproduserbarhet over dimensjonene tid, forskjellige kontekster og forskningsinstrumenter, og bemerker at innen den positivistiske forskningen er slik stabilitet en indikasjon på at en måling beskriver et sant fenomen heller enn en artefakt i forskningsprosessen(13). S*tabilitet* utgjøres også etter Hollenbæks (1978) definisjon gjentatte forsøks reproduserbarhet, mens *Riktighet* er graden av fravær av målefeil i forsøket(8).

Gwet (2019) fremlegger at *enighet* måles som en skårvariasjon mellom subjekter for kontinuerlige data, eller observeres som sammenfallende skår for kategoriske data, og at to granskere er enige når deres respektive skår er identiske for et gitt subjekt (41).Popping legger til at mens enighet observeres, er *reliabilitet* er de slutningene vi gjør ut ifra enigheten(8). Dersom enigheten måles over en terskel fastsatt av forskeren, kan resultatene ifølge Gwet omtales som reliable, og en kan anta at variasjoner i skårene utgår fra faktisk variasjon hos subjektene, og Popping (2019) tilfører at høy enighet indikerer at granskerne er utskiftbare(8, 41)Forfatteren legger til at den manglende transparensen av menneskelige kognitive prosesser kan redusere troverdigheten til forskning der høy enighet kreves(41).

Enighet telles som *n* subjekter der det er enighet for kategorien, og leses av (uthevet) i en krysstabell / enighetstabell. Total observert enighet leses av i nedre høyre celle.

Tabell 1 Enighet (Uthevet diagonal) med total observert enighet (Nedre høyre celle)

	Gransker 2		
Gransker 1	*Ja*	*Nei*	*Total*
Ja	**a**	b	a+b
Nei	c	**d**	c+d
Total	a+c	b+d	**N**

Dersom summen av de negative enighetene ikke er lik summen av de positive enighetene, er enighetene ujevnt fordelt (46). Dette som kan få konsekvenser for beregningen av IRR som diskuteres senere.

Krippendorf (2008) skiller ifølge Popping (2019) mellom *systematiske uenigheter*, som utviser en regularitet og til en viss grad kan forutsees, og *tilfeldige uenigheter* som er irregulære og ikke kan

[28] *Accu* uten feil s 2

forutsees(8), Systematiske feil kan ifølge Krippendorff stamme fra granskernes holdninger og preferanser, og kan de gi økt forekomst av type en feil eller inklusjonsfeil[29], det vil si at et karakteristika observeres hos subjektene observeres som ikke reelt er der. Tilfeldige feil er mer uforutsigbare feilklassifiseringer av raterne, og disse tendenserer til å øke forekomsten av eksklusjonsfeil[30] at et karakteristika eller en atferd ikke gjenkjennes(8)

Uenighet leses av som *n* i ikke- diagonale celler i enighetstabellen, som indikerer at granskerne har skåret forskjellige kategorier.

Tabell 2 Uthevet uenighet

	Gransker 2		
Gransker 1	*Ja*	*Nei*	*Total*
Ja	a	**b**	a+b
Nei	**c**	d	c+d
Total	a+c	b+d	N

3.4 Assosiasjon og intern konsistens

Innen kvalitativ forskning innhentes gjerne den samme informasjonen fra forskjellige subjekter ved spørreskjemaundersøkelser, og spørsmålssettet sees som reliabelt om det gir svært korrelert informasjon fra forskjellige respondenter på en konsistent måte(41). Gwet (2021) omtaler dette, itråd med Cronbach (1951) og Traubs (1994) definisjoner, som *intern konsistens*[31], og selv om høy intern konsistens ikke innebærer høy enighet mellom granskere, er den en indikasjon på at det er idemessig enighet mellom spørsmålene som stilles og beskrivelsene av det som skal undersøkes(41).

Der en gransker konsekvent gir lavere skår der den andre granskeren skårer høyere for det samme subjektet, kan ifølge Gwet og Popping den *rangerte ordenen* av skårene være gjennomgående lik og *konsistens* og *assosiasjon* [32]være høy, selv om det er stor forskjell på de faktiske skåringsverdiene og absolutt enighet er lav og omvendt (8, 41).

[29] Errors of commission, tilsvarer også type èn feil
[30] Errors of omission, tilsvarer også type to feil
[31] Beregnes oftest med Cronbachs Alpha, som er en annen reliabilitetskoeffisent (Gwet 2021)

[32] Assosiasjon måles med linjær regresjon Gwet (2019). Assosiasjon inngår ikke i analysene i denne oppgaven.

Figur 1 Høy assosiasjon og lav enighet eksempel

	Granskеr 2		
Gransker 1	*Ja*	*Nei*	*Total*
Ja	10	90	100
Nei	90	10	100
Total	100	100	200
(Prosent enighet =(10 +10)/200= 10%)			

3.5 Inter- rater reliabilitet og enighet

Reliabiliteten for granskeres skåringsprosses med et skjemaverktøy består av forsøkets *intra- og interrater reliabilitet*, og begge må beregnes for å kunne vurdere reliabiliteten eller validiteten av skjemaet(8) *Intra-rater reliabilitet* er reliabiliteten i gjentatte skåringer utført av den samme granskeren (47)[33], og beregninger av slik reliabilitet inngår ikke i denne oppgaven. Perfekt intra-rater reliabilitet er, ifølge Gwet (2021), imidlertid en forutsetning for å kunne vurdere Inter- rater reliabilitet,(41), og dette er en vesentlig begrensning for vår oppgave.

Interrater- reliabilitet er, Ifølge Zhao (2013), den mest brukte kvantitative indikatoren på målekvalitet i innholdsanalyse, og også mye brukt som kvalitetsindikator for diagnoseverktøy og tester(46). For målereliabilitet er inter- rater reliabilitet den bestanddelen som angir grad av samsvar i resultatene når samme fenomen måles eller vurderes av to uavhengige personer(48). Beregningen forutsetter, ifølge Carter og Lubinsky (2016), at utvalgene er like, målingene uavhengige og måletidspunktet det samme (48). Hanskamp- Sebregts et al. (2016) legger til at granskerne må bruke det samme instrumentet(3).

Popping (2019) baserer seg på Holstis (1969) definisjon, og omtaler *inter- rater reliabilitet* som en funksjon av granskernes evner, innsikt og erfaring, samt klarheten i kategoriene og graden av tvetydighet i dataene og rigiditet i kodingsprotollen(8). Forfatteren legger til at en kan se interrater-reliabiliteten som *konsistensen* i forsøket, det vil si at uavhengige granskernes skåringer har lite variabilitet fordi granskerne er utskiftbare og en har oppnådd konsistente forhold for begge skåringsprosessene. Popping oppsummerer at interrater- reliabiliteten er målbar som grad av likhet i konklusjoner for karakteristika ved et subjekt[34], og at dette representerer om dataene er korrekte representasjoner av variablene. Variablene trenger imidlertid ikke representere de sanne verdiene, og for å beholde validitetsperspektivet, kan en her legge til Gwets (2021) klargjøring som sier at om

[33] Intra- rater reliabilitet omtales ikke etter dette
[34] *Message or artifact* fritt omformulert og oversatt

to granskere har høy enighet, er skårene reliable, men om begge granskerne også er enige om den sanne skåren, er skåringene valide(41)

Etter definisjonene over kan vi forholde oss til *inter- rater reliabilitet* som variasjon mellom granskernes skår for samme subjekt, og *enighet* som de observerte antall målingene der granskerne skårer det samme subjektet i samme kategori. Vi tar med oss at i vår studie kan vi ikke si noe sikkert om validiteten av skåringene, og at en høy observert enighet eller høy reliabilitet også kan bety enighet om usann verdi, det vil si at granskerne utsettes for den samme målefeilen. For denne oppgaven beskrives inter-rater reliabilitet ved Interrater- reliabilitetskoeffisienter, det vil si et statistiske reliabilitetsmål som er et tall mellom -1 og 1. Reliabilitet diskuteres på et mer teoretisk nivå som variasjonen i skåringene, mens enighet- og uenighet brukes for å diskutere plasseringen av subjektene i de samme, henholdsvis forskjellige, kategorier. Når observert enighet eller observert uenighet omtales i oppgaven, er dette en telling av *n* i diagonalen henholdsvis ikke- diagonale celler i en enighetstabell.

3.6 Tilfeldig enighet

Spørsmålet om enighet mellom granskere er aktuelt, ifølge Mchugh (2021) fordi mennesker opplever og tolker fenomener forskjellig(30). Granskere vil, etter Poppings beskrivelser av uenighet, gjenkjenne karakteristika ved subjektene på forskjellige måter, og også tolke kategoridefinisjoner forskjellig(8). *Tilfeldig enighet* er ikke en falsk enighet, men fremstilles av Gwet (2019) som en type bonus som blåser opp det relative antallet subjekter der det er enighet uten at den kommer fra egenskaper ved skåringsmetoden(41).

Enigheten er tilfeldig om en av raterne skårer uavhengig av subjektenes karakteristika, eller ved å følge en ukjent bedømmelsesprosess uten åpenbar logisk sammenheng (41). Derfor mener Gwet, i likhet med Cohen (1960, at en skår som assosieres med tilfeldig enighet ikke er nyttig informasjon i reliabilitetsberegningen(41, 49), og at skårene heller ikke er identifiserbare(41).Om en forutsetter ren gjetning, vil det ved få kategorier være et begrenset antall muligheter for uenighet, mens det ved mange kategorier vil være lite sannsynlig at granskerne enes(8, 31).Et slikt tenkt scenario med ren gjetning beskriver et forsøks teoretiske *tilfeldige enighet* om det kun er antall kategorier som beskriver variablene som tenkes å ha betydning.

Etter blant annet Cohen (1960) og Bennet (1954) , er det uten kjennskap til den faktiske fordelingen av trekk i befolkningen kun den logiske sannsynligheten for tilfeldig enighet ut ifra antall kategorier som kan beregnes(31, 49), ved formelen

Figur 2 Utregning av tilfeldig enighet basert på antall kategorier

Tilfeldig enighet = 1/antall kategorier		
= 1/K		

(8, 49)

Enkelte kategorier kan kodes oftere enn andre på grunn av skjevhet i forekomst i befolkningen som studeres, eller som følge av at granskere vil identifisere forskjellige karakteristika ulikt, men på en konsekvent måte, og dermed ha en granskerspesifikk fordeling av subjektene over kategoriene(8). Hver granskers fordeling over kategoriene leses av i de ytre og nedre cellene, eller marginalene, i en enighetstabell og omtales som granskerens marginalfordeling.

Tabell 3 Granskernes marginaling og utregning av marginalfordeling

		Gransker 2		
	Gransker 1	*Ja*	*Nei*	*Total*
	Ja	a	b	**a+b**
	Nei	c	d	**c+d**
	Total	**a+c**	**b+d**	N
Eksempel marginalfordeling for svar "Ja" for gransker 1 = a+b/N				

Popping tilpasser Weisbergs (1974) teorier om forhold mellom variabler, og beskriver fire typer marginalfordeling. *Uniform marginalfordeling* innebærer at kategoriene har lik sannsynlighet for å brukes. *Marginal homogenitet*: at granskernes marginalfordeling er lik. *Marginal heterogenitet* innebærer at summmen av gangene en kategori brukes totalt delt på antall ratere, tilsvarer forekomsten av kategoriene. *Ingen marginal uniformitet*, innebærer at den ene granskerens skåringer ikke antydes av hvordan den andre skårer(8).

Feinstein og Cicchetti (1990) kritiserer begrunnelsen for marginalfordelingens rolle i reliabilitetsvurderinger, og mener at observatører som ikke er påvirket av bekreftelsesfeil vil tilnærme seg det som fremstår hos subjektet uavhengig av slike preferanser, forutsatt at de ikke kjenner til den større populasjonens fordeling over de samme kategoriene (50).Gwet legger til at antagelsen av en relativt fastsatt, granskerspesifikk marginalfordeling ikke er rimelig dersom granskerne er blinde for populasjonens fordeling(41).

Zhao(2013) går lengre enn de overnevnte forfatterne, og argumenter for at *tilfeldig enighet* i det hele tatt må se som en sammenlikningsreferanse og et teoretisk fenomen, men ikke som noe som faktisk inntreffer under skåringsprosesser ved vitenskapelig forskning(46). Popping bemerker at det kan være en grad av usikkerhet eller gjetning i skåringene uten at disse er rent tilfeldig fordelt, som at granskeren kan skåre enten kategori a eller b med en grad av gjetning dersom disse er relativt like, etter å ha utelukket kategori c(8). Her følges en ukjent begrunnelse, men granskeren vet at valget kan være feil, og dette er et problem som Popper observerer at i liten grad adresseres(8).

4.0 Beregning av Interrater- reliabilitet

Under presenteres de mål for interrater- reliabilitet som er valgt ut for oppgaven, begrenset til versjonene som er aktuelle for dikotomt målenivå[35]. Til slutt beskrives tolkning av IRR-koeffisientene og det begrunnes kort hvorfor andre mye brukte koeffisienter er valgt bort.

4.1 Prosent enighet

Prosent enighet[36] er et gammelt, intuitivt og mye brukt mål på interrater- reliabilitet som regnes ut som antall enigheter delt på totalt antall skåringer:

Tabell 4 Utregning av prosent enighet

		Gransker 2		
	Gransker	*Ja*	*Nei*	*Total*
	Ja	a	b	a+b
	Nei	c	d	c+d
	Total	a+c	b+d	N
% enighet = Alle enigheter/ alle skåringer				
= (a+d)/N				

Prosent enighet bygger på et enighetskonsept som antar at all *skåring er velinformert*[37], og interrater- reliabilitetsmålet som utgår fra dette forutsetter at det aldri oppstår tilfeldig enighet, noe som ble kritisert allerede i 1960 av Cohen da han lanserte sin koeffisient som korrigerte for tilfeldig enighet(46, 49). Paradokset som følger av dette enighetskonseptet, er at *tilfeldig gjetning kan være reliabelt*, også i virkelige *situasjoner*, der enkelte subjekter er vanskelige å skåre og det er en grad av usikkerhet[38]*(46)*. Dette paradokset inntreffer også der skåringsoppgaven er så vanskelig at det i praksis ikke er annet enn tilfeldig gjetning, og det blir tydeligst ved dikotome variabler der både

[35] Prosent enighet og Cohens Kappa finnes også for ordinalt nivå / som vektede versjoner, disse omtales ikke i oppgaven.
[36] Omtales også som "raw agreement"
[37] Basic assumption 1 i Zhao, på engelsk brukes *Honest*
[38] Paradox

prosent enighet og forventet tilfeldig enighet[39] blir 50%, dvs midt mellom ingen og perfekt reliabilitet(46). Gwet stiller spørsmål ved om prosent enighet og tilfeldig enighet har en tilstrekkelig felles matematisk grunnlag til at forskjellen mellom dem gir mening(41), og Popping fremhever at ettersom prosent enighet ikke korrigerer for tilfeldig enighet, kan målet påvirkes forskjellig av granskernes marginalfordeling ved samme enighet, noe som er problematisk for sammenlikninger med prosent enighet for forskjellige situasjoner(8)

Allerede i 1954 ser en Bennett, blomquist and goldstein bemerke hvordan tolkningen av prosent enighet[40] må begrenses ved en nedre cutoff- verdi, beregnet ved utregnet tilfeldig enighet(31). Dette betyr at beregnet prosent enighet i virkeligheten ikke går fra 0 til 1, men fra *forventet tilfeldig enighet* ved 1/k og til 1 som innebærer perfekt enighet.

Prosent enighets tendens til å overberegne enighet er bemerket av svært mange forfattere (30, 31, 41, 49), mens Hallgren (2012) påpeker at nær perfekt enighet kan aksepteres for høyere målenivåer, og at prosent enighet bør oppgis som supplement til mål som korrigerer for tilfeldig enighet(27) Zhao oppsummerer at, selv om prosent enighet er beskrevet av Cohen (1969) som den *mest primitive* og av Hayes og Krippendorff (2007) som den *mest feilaktige* reliabilitetsindikatoren, er prosent enighet ikke et dårlig interrater- reliabilitetsmål, men best egnet der velinformert skåring og ingen tilfeldig enighet kan antas fordi en har en godt utviklerotokoll og subjektene er enkle å klassifisere(46).

4.2 Korrigering for tilfeldig enighet

Mens prosent enighet utgår fra at all skåring er oppriktig. finnes en mengde[41] koeffisienter som korrigerer for tilfeldig enighet og beregner et desimaltall mellom -1 og 1 som mål for IRR. Cohens Kappa som etter Zhao, Liu og Dengs klassifisering baserer seg på *tilfeldig enighet som en funksjon av observert enighet*, og Gwets AC1 som utgår fra *både antall kategorier og observert fordeling(46).* Det finnes en mengde koeffisienter som korrigerer for tilfeldig enighet, og mange av disse tilsvarer hverandre matematisk. Zhao et al (2013) identifiserte 11 matematisk unike koeffisienter for kategoriske data, og diskuterer i sin artikkel Pardokser og abnormaliteter ved disse. Under presenteres koeffisientene som er valgt ut for oppgaven,

[39] (1/k= 1/2) ved dikotome variabler
[40] Bennett, Blomquist and Goldstein kalte prosent enighe «Raw stability proportion».
[41] Zhao Liu og Deng identifiserer 11 matematisk forskjellige Koeffisienter.

4.2.1 Tilfeldighetskorrigerte mål: Cohens Kappa

Cohens Kappa ble presentert i 1960 som et interrater- reliabilitetsmål som kunne erstatte prosent enighet ved å korrigere observert enighet for tilfeldig enighet(49). Kappa er idag det mest brukte tilfeldighetskorrigerte interrater-reliabilitetsmålet for nominale datatyper med to granskere42 (41, 46, 51), og beregnes på denne måten etter Gwet :

Figur 3 Cohens Kappa utrenging

$K=(P_o-P_e) / (1 - P_e)$
Cohens Kappa= Observert enighet- Tilfeldig enighet/1-Tilfeldig enighet
Eller
1-(1-Observert enighet)/(1-Tilfeldig enighet)

K= Observert enighet- Tilfeldig enighet/ 1- Tilfeldig enighet

Det at Cohens kappa beregner tilfeldig enighet som en *funksjon av observert fordeling*, gjør dette målet svært avhengig av at marginalene er balanserte eller *homogene*, det vil si at hver kategori har omtrent like stor sannsynlighet for å brukes (41, 46). Denne fordelingen beregnes for kappa per gransker, under antagelsen av at forskjellige granskere vil ha forskjellige tendenser i sine skår. De tilfeldige klassifiseringene beregnes etter marginal sannsynlighet, og denne nærmer seg henholdsvis 1 og 0 der marginalene er ubalanserte, vil utregningen av tilfeldig enighet ifølge Gwet (2019) bli urimelig stor(41).

Feinsten og Cicchetti (1990) beskriver at kappas avhengighet av marginal homogenitet er *granskerspesifikk*, heller enn s*ubjektspesifikk*, og ikke samsvarer med den fordelingen som ved vanlige forsøk vil utgår fra karakteristika ved subjektene som skåres(50). Gwet (2021) og Zhao, Liu og Deng (2013) beskriver begge at Kappa utgår fra at granskernes fordelinger er statistisk uavhengige av hverandre, noe de ikke kan være om granskerne skårer det samme utvalget på bakgrunn av subjektenes karakteristika(41). Gwet (2019 fotnote s. 21) formulerer den samme tanken som at skåringene er uavhengige dersom en ut ifra kunnskapen om den enes skår ikke kan forutse den andres skår(41), og dette er en definisjon som en kan påstå fører til at en observert

[42] Det finnes også videreutviklede kappatyper som PABAK (prevalence adjusted bias adjusted kappa) og vektet kappa for ordinalt målenivå.

enighet som er høyere enn forventet tilfeldig enighet utelukker statistisk uavhengighet. Popping (2016) legger til at kappas enighet er svært snevert definert, som de *tilfeldige uenighetene som ikke inntreffer*(8). Zhao, Liu og Deng (2013) går enda lengre, og presenterer matematiske argumenter for å begrunne at Kappa faktisk utgår fra at all koding er tilfeldig(46)

Kappa listes av Zhao, Liu og Deng som et mer konservativt IRR- mål enn Gwets AC, som i vanlige situasjoner vil ligge under Gwets AC og Prosent enighet. Den utbredte bruken av Kappa er ifølge flere forfattere vanskelig å forsvare, men en forklarng kan være at forskere gjerne tolker konservative beregninger som rigiditet i forskningsprotokollene og også tendenserer til å være mer skeptiske til type 1 enn type 2 feil(46). Forfatterne påpeker at utstrakt bruk av et konservativt IRR-mål og det at det er mer sannsynlig at forskere publiserer dersom IRR er god, kan ha hindret publikasjon av mer skjeve fordelinger, og fått verden til å se mer normalfordelt ut(46).

4.2.2 Tilfeldighetskorrigerte mål: Gwets AC1

I boken *Handbook of inter- rater reliability* redegjør den amerikanske statistiskeren Kilem L. Gwet for utfordringer ved etablerte metoder for beregning av interrater- reliabilitet, og presenterer så Gwets AC1 [43] som et kappaparadoksresistent alternativ. Gwets AC1 er ved en metodesammenlikningsstudie av Wongpakaran (2013) funnet å produsere mer stabil IRR og å være langt mindre påvirket av forekomst og marginalfordeling enn Cohens Kappa samt ligge nær prosent enighet(52). Gwets AC1 beregnes ved formelen:

Tabell 5 Gwets AC1 utregning

$AC1 = p\text{-tilfeldig enighet} / 1\text{-tilfeldig enighet},$
$p = \text{Enighetsforholdet} = N(A_{Ja}, B_{Nei}) / N(\text{Totalt}),$
$\text{Tilfeldig enighet} = 2q(1 - q),$
$q = N(A_{Ja}) + N(B_{Ja}) / N(\text{Totalt})$

Ifølge Zhao, Liu og Deng (2013), er denne koeffisienten unik, ettersom den baserer tilfeldig enighet både på antall kategorier og fordeling[44](46). Gwets Ac tar utgangspunkt i at det kun vil være et mindretall av observasjonene som vil gi tilfeldig enighet mellom granskerne(46). Enighetskonseptet baseres derfor på at utvalget gjøres fra en «nedtrimmet» befolkning, der de de subjektene som klassifiseres uavhengig av subjektkarakteristika først ekskluderes (41). I dette utvalget er tilfeldig

[43] Også kalt *Gwets Gamma*

enighet en umulighet, og den beregnede koeffisienten representerer den enigheten granskerne oppnår *med hensikt*, uten de *vanskelig klassifiserbare* subjekter som gir tilfeldig enighet(41).Granskerfordelingen bygger på et gjennomsnitt av skår for begge granskere per kategori[45], og antar dermed at granskerne ved objektiv skåring av de samme subjektene vil påvirkes av karakteristika hos disse til å skåre ut ifra en sammenslått sannsynlighet(41, 46).

Ifølge Zhao, Liu og Deng ligger Gwets enighetskonsept nærmere et reelt skåringsscenario enn andre koeffisienter som Cohens kappa, men adopterer også antakelser som ved vanlige forskningssituasjoner brytes og gir paradokser.

4.2.3 Antagelser for Cohens Kappa og Gwets AC1

Zhao Liu, og Deng (2013) identifiserte flere paradokser[46] som oppstår der antagelsene til grunn for IRR- koeffisientene ikke oppfylles. Cohens Kappa og Gwets AC deler 6 av paradoksene mens Cohens Kappa rammes av ytterligere 9 paradokser og Gwets Ac av 3(46). Under presenteres paradoksene for vårt koeffisientutvalg, begrenset til de som er aktuelle for en dikotome variabler med tilstrekkelig store utvalg.

Felles for alle tilfeldighetskorrigerte interrater- reliabilitetsmål er, ifølge Zhao, Liu og Deng, antagelsen om at granskere vil *maksimere tilfeldig koding* og begrense oppriktig skåring til de tilfellene som gjenstår etter at mulighetene for tilfeldig koding er benyttet. Forfatterne illustrerer dette ved et teoretisk scenario der granskerne som skal skåre et subjekt først trekker klinkekuler fra en urne med et fastsatt antall klinkekuler av forskjellige farger. Dersom de trekker like klinkekuler, skåres dette som en *tilfeldig enighet*, og bare ved *tilfeldig uenighet* og ulike farger ser de på subjektet og skårer oppriktig. Dette innebærer at tilfeldig koding går foran, tillater begrenser og innnehar all ærlig koding[47], etter antagelser om *maksimering av tilfeldighet og begrenset oppriktighet*. Ifølge forfatterne opererer ikke granskere etter et slikt scenario i praktiske forsøk, og dersom systematisk tilfeldig skåring inntreffer på denne måten, skal dataene skal kastes(46).

Artikkelforfatterne refererer videre til Riffe (2005) og Grove et al. (1981), som beskrev at selv om enighet *kan* inntreffe tilfeldig, kan enighet også være resultatet av en velutviklet protokoll eller subjekter som er lette å skåre(46). Granskere vil, ifølge Zhao. Liu og Deng, forsøke å skåre riktig, og normalt skåre lette subjekter med små eller ingen feil, og resterende ved å tillate en grad av gjetning eller tilfeldighet[48]. Dette innebærer en antagelse om *variabel tilfeldighet og total*

[45] Conspired quota fritt oversatt
[46] ZHao, Liu og Deng (2013) omtalte *Paradokser og abnormaliteter,* men skillet mellom disse ser ut til å ikke være relevant for oppgaven.
[47] Zhao, Liu og Deng (2013) benevner dette et Goodman- Guttmann scenario, etter Guttman 1946 og Goodman og Kruskal (1954)
[48] Zhao, Liu og Deng (2013) beskriver dette som et Grove- RIffe- Scenario.

oppriktighet, og det finnes ifølge Zhao, Liu og Deng ingen koeffisienter som baserer seg på dette scenariet, (46)

Forfatterne bemerker at ettersom Gwets AC skiller vanskelige subjekter fra lette og dermed nærmer seg *total oppriktighet*, gir ikke denne koeffisienten like mange paradokser som kappa, men *maksimum tilfeldighet- antagelsen* gjør at heller ikke Gwets AC bygger på et reelt skåringsscenario(46). Oppsummert utgår både Kappa og Gwets AC utgår fra at tilfeldig skåring er reell, at denne maksimeres av granskerne på forskjellige måter og må korrigeres for, og at ærlig skåringer begrenser seg til en andel av subjektene som er begrenset og definert av tilfeldig skåring(46). Begge koeffisienter baserer seg dessuten på hver sine definisjoner av tilfeldighet, og det er, ifølge Zhao, Liu og Deng, vanskelig å begrunne valget for disse.

4.2.4 Paradokser

Variasjoner av antagelsene om *maksimering av tilfeldighet og begrenset oppriktighet* gir paradokser og abnormaliteter som er oppsummert i tabeller i vedlegg. Under og de følgende kapitlene presenteres de paradoksene som er mest relevante for vårt skåringsscenario etter forenklede tabeller fra Zhao, Liu og Deng

4.2.5 Felles paradokser for Cohens Kappa og Gwets AC

Maksimering av tilfeldighet og begrenset enighets- antagelsene gir, ifølge Zhao, Liu og Deng grunnleggende paradokser i utregningene av både Kappa og Gwets AC.

Tabell 6 Felles paradokser for Cohens Kappa og Gwets AC1

Nr.	Paradoks/ Abnormalitet
2	Ingenting annet enn tilfeldighet
3	Epler kan sammenliknes med appelsiner
4	Mennesker er en undergruppe av menn
5	Pandaer er en undergruppe av menn
19	Terskelen flytter seg
20	Sirkulær logikk

Det demonstres ikke hvilke konkrete utfall dette får for IRR, men slike konseptuelle beskrivelser kan synes å gi bedre forståelse hos ikke- statistikere av at antagelsene bak koeffisientene kan være grunnleggende gale og gi store feilutslag.

Likninger som er felles for begge koeffisienter gjør at *ærlige enigheter* sammeliknes med *tilfeldige uenigheter*, at en regner ut hvilken andel *observerte uenigheter* utgjør av *tilfeldig uenighet*, og også

at *ærlige enigheter* og *observerte uenigheter* er to undergrupper av tilfeldige uenigheter(46). Forfatterne beskriver dette som *at epler sammenliknes med appelsiner, mennesker er en undergruppe av menn,* og *pandaer og mennesker er en undergruppe av menn* (46). Det teoretiske skåringsscenariet som maksimerer tilfeldig enighet, sier også at ettersom tilfeldig uenighet i klinkekuletreningen innebærer alle oppriktige skåringer, utgjør tilfeldig enighet og tilfeldig uenighet alle skåringer som finnes, noe forfatterne omtaler som *ingenting annet enn tilfeldighets- paradokset*.

Alle koeffisienter som i en viss grad baserer seg på fordeling, er ifølge Zhao, Liu og Deng, sensitive for fordelingsskjevhet(46). Svært skjev fordeling vil produserer en høy tilfeldig enighet, som er terskelen observert enighet må over for å gi en kappa som er over 0, og som også regnes inn i Gwets AC, i tillegg til antall kategorier. Det at tilfeldig enighet er en *terskel som flytter seg* med fordelingen, mener forfatterne at ikke er en rimelig avhengighet for IRR- mål(46).

Et paradoks som ser ut til å være svært aktuelt for vår studie, er *sirkulær logikk*- paradokset, der IRR målene skal undersøke reliabilitet ved et konstrukt eller instrument, men selv også påvirkes av elementer, som semantiske faktorer eller kategorier, ettersom disse påvirker fordeling og også bestemmer antall kategorier(46).

4.2.6 Særegne paradokser for Cohens Kappa

Det påpekes av svært mange forfattere at Kappas sjansekorreksjon kan gi *kappaparadokser*, de hyppigst nevnte er at skjevhet i fordelingene vil omgjøre en relativt høyere prosent enighet til en relativt lavere K (8, 41, 46, 50).

Tabell 7 Særegne paradokser ved Cohens Kappa (46)

Nr.	Antagelser og paradokser Cohens kappa
10	Høy enighet og lav reliabilitet
11	Udefinert reliablitet
12	Ingen endring i observert enighet, stort fall i r
13	Ingen uenighet, ingen økning i reliabilitet
14	Lite økning i enighet, stor økning i reliabilitet
15	Økt observ. Enighet, stort fall i r
16	Oppriktig koding tilsvarer myntkast
17	Bedret koding straffer seg
18	Enighet straffer seg

Ettersom en skjev fordeling gir høyere tilfeldig enighet og mindre rom for oppriktig skåring, vil høy tilfeldig enighet gi en høy terskel for tilfeldig enighet der fordelingen er skjev. Dette omtales av

Zhao, Liu og Deng (2013) som paradoksene *Enighet straffer seg og uenighet belønnes og Bedret koding straffer seg Ærlig koding som gir bedre sensitivitet og spesifisitet kan være like dårlig som å kaste en mynt.*

Forfatterne demonstrerer dette ved et tenkt forsøk der 50 subjekter i et utvalg på 60 har et karaktertrekk som er vanskelig identifiserbart. Dersom èn gransker finner karaktertrekket hos alle 60 og har 10 falsk positive mens den andre finner trekket hos 40 og har 10 falsk negative, gir de 40 enighetene en prosent enighet på 66,7% men en kappa på 0, som er langt under enighet ved tilfeldig myntkast på 50%..

Dersom to granskere enes på at fordelingen over to kategorier er 100% og 0%, kan ikke Kappa beregnes og gir *udefinert reliabilitet*, selv om enigheten er100%, Her har begge granskere i et kappa- scenario kun klinkekuler av den samme fargen i sin urne, og det at en trekning alltid vil gi enighet utelukker oppriktig koding(46). Zhao, Lui og Deng kritiserer Krippendorffs (2004) som forsvarte dette paradokset med at det uten variasjon i dataene heller ikke kan være reliabilitet. Forfatterne argumenterer for at ettersom alle subjektene kan inneha det samme karakteristika, bør reiabiliteten være god og kunne beregnes, under forutsetningen at granskerne skårer oppriktig (46).

Kappa utgår ifølge Zhao, Liu og Deng, fra en individuell fordeling av skår over kategoriene per gransker, noe som innebærer at i et forsøk der 1000 subjekter skåres på to kategorier og granskerfordelingen er 999/1 og 998/2, blir Kappa negativ og indikerer lavere enighet enn ved tilfeldig trekning. Dette er fordi kappa maksimerer tilfeldig enighet på en måte som kan beskrives ved at hver gransker fyller sin urne med tilsvarende forhold farger på klinkekulene som det som observeres ved skåring. De to gangene klinkekulene er forskjellige er her de eneste mulighetene for å skåre oppriktig, og for begge disse er det uenighet noe som gir negativ Kappa. Dette omtales som *Høy enighet, lav reliabilitet paradokset* av flere forfattere(8, 41, 46, 50)

Flere paradokser der kappa enten faller eller stiger dramatisk ved en liten ingen endring i observert enighet, oppstår ifølge Zhao, Liu og Deng (2013) ved ekstreme fordelinger, noe som stammer fra det samme prinsippet om individuell fordeling(46). Paradoksene kan ta formene:

Ingen endring i observert enighet, Stort fall i K,
Liten økning i enighet, Stor økning i K
Økt observert enighet, stort fall i K,
Ingen uenighet, ingen økning i K
Samme observerte enighet, stor endring i K

Disse paradoksene antas å være vanskelig å observere i vårt forsøk der det kun er ettskåringstidspunkt, og heller ikke en såpass ekstrem skjevfordeling av dataene som der paradoksene demonstreres av Zhao, Liu og Deng. En må imidlertid være klar over at paradoksene kan være til stede i forskjellig grad ettersom vi forventer en grad av skjevfordeling på bakgrunn av forekomst av helsehjelpsproblemer i andre studier,

4.2.7 Særegne paradokser for Gwets AC1

Gwets AC1 produserer ifølge Zhao, Liu og Deng langt færre og mindre dramatiske paradokser enn Kappa, men i forsøket på å kompensere for kappaparadoksene introduseres oppstår det andre paradokser som Kappa ikke rammes av(46).

Tabell 8 Gwet AC1s særegne paradokser

	Gwets AC særegne paradokser og abnormaliteter
6	Økende antall kategorier øker reliabilitet
21	Samme kvalitet og observerte enighet, høyere r
22	lavere kvalitet og bservert enighet, høyere r

For å motvirke effekten av at flere kategorier innebærer flere kulefarger og flere muligheter for ulike farger og dermed tilfeldig koding, regnes antall kategorier inn i tilfeldig enighetsberegningen for Gwets AC1 ved å gange med 1/(k-1), i tillegg til fordeling(41, 46). Dette senker tilfeldig enighet, og gir også en annen implikasjon, nemlig at *Økende antall kategorier øker reliabilitet*, også for tomme kategorier(46). For vårt forsøk med dikotome variabler er ikke dette en aktuell problemstilling ettersom 1/ (2-1) = 1, men anerkjennelsen av problemet gjør det desto viktigere å håndtere i utgangspunktet dikotome variabler med en ytterligere «vet ikke» kategori , som er en type kategori i vårt variabelutvalg.

Gwets AC antar at granskerne enes om et forhold av klinkekuler i urnen[49] før de trekker ettersom måldistribusjonen går fra å være jevn til ujevn innebærer dette ujevn klinkekule, mindre mulighet for ulike klinkekuler, mindre tilfeldig enighet, en lavere terskel og høyere ac. grunnen til at dette ikke er normale utfall, er at det heller ikke er på denne måten granskere enes.

Det at Gwets AC utgår fra at granskerne går sammen om en fordeling av klinkekuler i urnene før trekning er, ifølge Zhao, Liu og Deng heller ikke en antagelse som følger det virkelige, *variabel tilfeldighet og fullstendig oppriktighets*- scenariet. Dette fører til at K som et unntak blir høyere enn

[49] *Conspired quota* kalles dette av Zhao, Liu og Deng (2013)

Gwets AC dersom tilfeldig enighet blir svært lav fordi granskerne har svært ujevne fordelinger i hver sin retning(46).

Dersom 100 subjekter skåres på to kategorier med fordelingen enighet for 40 negative og 40 positive, og uenighet for 10 positive og 10 negative, gir dette observert enighet 80% men Gwets Ac er 0,6. Dersom 40 av de positive enighetene blir negative enigheter, er enigheten den samme, men ACc hopper til 0, 75. Dersom fordelingene gjøres ujevne ved at det blir uenighet for 4 av de 80 subjektene det tidligere var enighet for, innebærer dette en vanskeligere oppgave og lavere skåringskvalitet, men mens prosent enighet er 76% er Gwets AC 0, 7 og høyere enn den opprinnelige 0, 6. Disse situasjonene omtales som *Samme kvalitet og observerte enighet, høyere IRR Lavere kvalitet og observert enighet,* høyere IRR paradoksene(46),

4.3.8 Generelle observasjoner

Det siden 1950 tallet vært enighet om at prosent enighet er et utilstrekkelig mål på enighet, fordi det ikke korrigerer for tilfeldig enighet. Den allmenne antagelsen om at tilfeldig enighet må korrigeres for modereres imidlertid noe av Zhao, Liu og Deng (2013), selv om viktigheten av å beregne tilfeldig enighet også vektlegges (46). En antagelse om ingen tilfeldighet som følger prosent enighet vil overestimere oppriktig enighet, mens antaglesen om *maksimering av tilfeldighet* kan underestimere eller overestimere IRR, uten at vi vet et ikke når, hvor mye eller i hvilken retning.

Fra spesielt Zhao, Liu og Dengs argumentasjon tar vi med oss videre at ingen koeffisienter per i dag baseres på et reelt skåringsscenario(8). Gwets AC baserer seg både på fordeling og kategorier og er nærmere antagelsen om fullstendig oppriktighet, noe som gjør den til den koeffisienten som ligger nærmest et *variabel tilfeldighet og total oppriktighets- scenari(8)o*. Cohens Kappa er svært mye brukt i helseforskningen, selv om det er allment kjent at det følger paradokser med målet i svært mange situasjoner.

Gwets AC påpekes av Zhao, Liu og Deng å være generelt et liberalt mål(46), igjen uten at det beskrives hvilken referanseverdi dette er i sammenlikning med. Det oppfordrs til forsiktighet der en ser høy IRR beregnet ved liberale mål, og forfatterne problematiserer også det at forskere kan «shoppe rundt» etter en koeffisient som gir høy IRR(46). Det påpekes imidlertid at Kappas tendens til å beregne en konservativ IRR, også appellere til forskeres tendens til å tolke lave estimater som et tegn på rigiditet i forskningen(46). Det at kappa imidlertid ser ut til å kun belønner økt enighet for relativt normalfordelte data, mener Zhao, Liu og Deng gjør målet upassende i mange praktiske forskningssituasjoner(46).

4.3 Tolkning av IRR-koeffisienten

Cohen (1960) beskriver Kappa som *enighet utover tilfeldig enighet i prosent, uttrykket som et desimaltall(49),* og konseptuelt ser dette også ut til gjelde for Gwets AC. I vår litteraturutvalg presenterer forfatterne gjennomgående desimaltallet for IRR- koeffisentene uten å gjøre om dette til prosent, og velger i tillegg gjerne en kvalitativ tolkning av IRR etter kjente kriterier. Popping observerer at det å gi inntrykk av at reliabilitet kan beskrives ved et tall har en villedende appell ettersom reliabilitet også har ikke- lineære karakteristika. Det kan være langt vanskeligere å øke reliabilitet fra .90 til .95 enn fra .50 til .55, noe som kan gjøre en linjær tolkning upassende(8).

Det bemerkes av Gwet (2019) at det at Kappa korrigerer for forventet tilfeldig enighet, gjør Kappa til et relativt mål(41), og dette er en bemerkning som må kunne generaliseres til alle IRR-koeffisientene. Feilestimat i form av konfidensintervall for skal også oppgis for reliabilitetsberegning i tråd med anbefalinger fra flere forfattere (8, 27, 41, 47). En konservativ tolkning etter Popping(2019), er at den nedre konfidensintervallgrensen skal overstige en reliabilitet som er forhåndsdefinert som akseptabel for forskeren, og dersom selve IRR målet ikke når den forhåndsdefinerte grensen, skal ikke dataene analyseres videre(8).

IRR- koeffisientene gir en beregning av variasjon eller enighet ved et desimaltall mellom -1 og 1(8, 30, 46). Det er enighet blant forfattere om at en IRR på 1 innebærer perfekt enighet Ved IRR på 0 gir skåringene en *enighet som ved tilfeldig skåring* etter Poppings definisjon(8), mens Zhao, Liu og Deng beskriver dette som *ingen reliabilitet*, og Ten Hove et. al som *ingen enighet*(51)[50]. Negativ IRR innebærer dårligere enighet enn ved tilfeldig skåring(8, 46).

Forutsatt at det er variasjon i dataene, vil alle IRR- koeffisienter bli 1 der prosent enighet er 100, og det er først der prosent enighet faller fra 100 at forskjeller mellom koeffisientene kan oppstå(46).

Generelt er IRR – Koeffisienter funnet å produsere svært forskjellig IRR for de samme dataene(8), og det er ikke enighet om noen grenseverdi for god nok reliabilitet eller intervaller for kvalitative tolkninger. De overveldende fleste forfattere presenterer resultater fra reliabilitetsstudier både ved koeffisienten og den kvalitative tolkningen, og dette er i tråd med Kottners (2011) og Hallgrens (2012) anbefalninger. Desidert mest brukt er de kvalitative tolkningene fra Landis og Koch (1977), mens andre tolkninger er foreslått av Cohen (1960) og Altmann (1991).

[50] Den sistnevnte definisjonen av Ten Hove er muligens ikke rimelig da en tenker seg at negativ IRR må innebære dårligere enighet enn ingen enighet

Tabell 9 Kvalitative tolkninger av IRR

Cohen (1960)		Altmann (1991)		Landis og Koch (1977)	
IRR	Tolkning	IRR	Tolkning	IRR	Tolkning
≤ .00	No agreement			<.00	Poor
.01-.20	None to slight	<.20	Poor	≤ .20	Slignt
.21-.40	Fair	.21-.40	Fair	.20-.40	Fair
.41-.60	Moderate	.41-.60	Moderate	.41-.60	Moderate
.61-.80	Substantial	.61-.80	Good	.61-.80	Substantial
.81-1	Almost perfect	.81-1	Very Good	.81-1	Almost perfect

(49, 53, 54)

Ten Hove (2021) undersøkte hvordan valg av IRR påvirker tolkningen ved landis og kochs anbefalninger, og fant at forskjellige koeffisienter varierte mellom dårlig (<0,21) til nesten perfekt (>0,8) for de samme dataene, noe som gjorde at forfatterne stilte spørsmål ved verdien av beregning av IRR i det hele tatt(51). Mchugh påpeker at det at Landis og Kochs grense for akseptabel IRR ved 0,41 innebærer at det at 40% av skåringene kan være feil Mange anbefaler 80% enighet.

Krippendorff anbefaler en mer konservativ tolkning enn Landis og Koch, og mener at det ikke kan gjøres konklusjoner på bakgrunn av dataene om IRR er mindre enn 0,67, mens disse kan gjøres med forbehold mellom IRR 0,67 og 0,8, og med sikkerhet over 0,8. Det ser ut til å være enighet hos flere forfattere at en IRR over 0,8 er god nok. Krippendorff baserer dette på innholdsanalyse og annerkjenner at akseptabel IRR kan avhenge av studien og forskningsspørsmålet.

Kanskje kan en argumentere for at journalgjennomgang har et element av innholdsanalyse ettersom det er kvalitative tekster og konsepter som skal vurderes og måles, og at Krippendorffs grenseverdi på 0,8 er passende ettersom det ellers ikke på forhånd er fastsatt noen akseptabel reliabilitet for vårt forsøk. I alle fall kan en være enig om at en IRR ned mot 0, 41 som kan se akseptabel ut etter Landis og Kochs kriterier, vil være langt fra god nok i vårt forsøk. Ettersom det i beskrivelsen av variabler som forebyggbart dødsfall er iberegnet mye usikkerhet, bør vi kanskje også velge en konservativ tilnærming uten rom for feilmargin og ønske oss nedre konfidensintervall over 0,8.

Det at IRR 0,8 kan være god nok reliabilitet innebærer ikke, etter det vi har gjennomgått, at prosent enighet på 80% er tilstrekkelig enighet. Av de foregående kapitlene vet vi at skjevhet i dataene øker muligheten for tilfeldig enighet og at få kategorier gjør det samme, og ettersom vi forventer skjevhet og har dikotome variabler, der sannsynligheten for tilfeldig enighet uten å regne inn

fordeling allerede er 50%, er det usikkert hvilken grense en skal sette for et mål for IRR som ikke er korrigert for tilfeldig enighet. Det at vi anser at vi har en rigid protokoll og et utvalg der de fleste subjektene vil være relativt lette å kategorisere av granskere som er godt vant til slike vurderinger, er argumenter imot at vi vil ha mye tilfeldig enighet, selv om vi kan anta at vi ville hatt mindre tilfeldig enighet ved flere kategorier og mer normalfordelte variabler.

Dersom en ser har stor skjevhet i granskermarginalene tendenserer Gwets AC til å være urimelig høy(46), mens Cohens kappa beregnes enten høyt eller lavt(41, 46). Her ser prosent enighet ut til å være et mer rimelig estimat ifølge Zhao, Liu og Deng (Tabell 19.11)(46), men referansemetoden som er benyttet for å definere hva som er urimelig høyt eller passende oppgis ikke.

Utover tolkningen av IRR, fremlegger Gwet (2019) skal en for overføringsverdien av IRR fra reliabilitetsstudier vurdere om en kan anta at andre granskere vil enes for det samme utvalget og at de samme granskerne vil opprettholde enigheten om de skårer andre subjekter og på bakrunn av dette konkludere med om enigheten «hører til» dette ene eksperimentet når akkurat disse granskerne skårer disse subjektene. Gwet påpeker at konklunsjonen av denne vurderingen svært ofte vil vær at overføringsverdien er lav, ettersom forutsetningene om at granskerne og subjektene er gode representanter for sine populasjonen og at protokollen har høy grad av rigiditet sjelden er tilstede samtidig (41)

5.0 Metode

5.1 Design

Hovedstudien er en retrospektiv observasjonsstudie. Oppgaven inngår i en pilotstudie for å vurdere alternative skåringsinstrumenters egnethet for journalgjennomgang i hovedstudien, og er en fullt krysset interrater- reliabilitetsstudie.

5.2 Utvalg

Et utvalg av 200 journaler fra i alt 1081 dødsfall i 2014 ved OUS ble gjort via en nettjeneste som genererer tilfeldige utvalg. Alle journaler der dødsfall inntraff under sykehusoppholdet ble inkludert, og ingen eksklusjonskriterier er oppgitt. Utvalget utgjør 18,5% av populasjonen og utvalgsstørrelsen vurderes å være tilstrekkelig stort etter blant andre Kraemer (2002) (kap. 2.5.1), og et relativt stort utvalg for interrater- reliabilitetsstudier etter anbefalingene fra Håndbok for Journalgjennomgang(11), og er vurdert å være *representativt* ettersom det er tilfeldig og av tilstrekkelig størrelse for pasientene som døde ved sykehuset i 2014. Utvalget antas også å være *tidskonsistent*, ettersom det ikke foreligger vesentlige endringer i utskrivelsespraksis ved OUS siden 2014.

5.2 Granskerutvalg

Granskerne ble valgt ut blant legespesialister ved Oslo universitetssykehus etter kriteriene *lang klinisk erfaring*, *erfaring med journalgjennomgang*, samt *interesse og tilgjengelighet.*

Utvalget må sies å være et beleilighetsutvalg, og det er ikke undersøkt om disse er representative for befolkningen av legespesialistene ved OUS. Dette begrenser utskiftbarheten av granskerne og til tross for at uavhengighet for gjennomgangene er sikret

Demografiske karakteristika samt erfaring og utdanning for legegranskerne er listet i tabell under:

Tabell 10 Egenskaper ved legegranskerne

	Legegransker A	Legegransker B
Alder	62år	60år
Kjønn	Kvinne	Mann
Spesialisering	Generell og ortopedisk kirurgi	Geriatri
Klinisk erfaring	38 år	36 år
Nåværende stilling	Spesialrådgiver forskning, 20% klinisk arbeid	Prof. II, Overlege
Forskningserfaring	ja	ja
Erfaring med journalgjennomgang	ja	ja, omfattende

Under antagelsen av at granskerens kjønn har liten eller ingen påvirkning på skåringene, bemerkes det at våre legegranskere er svært sammenliknbare for egenskapene alder, klinisk erfaring og forskningserfaring og dermed egnet for beregnet av interrater- reliabilitet.

Granskerne har forskjellige spesialiteter, noe som kan tenkes å påvirke variasjonen dem imellom, men PRISM2 tar høyde for dette ved at det inngår en egenvurdering av hvorvidt mangel på subspesialitet har hemmet granskernes evne til å skåre subjektene, og det er tilrettelagt for å konsultere med andre spesialister for avklaring av dokumentasjonen[51]. Legegransker A benytter

Utskiftbarheten av våre granskere kan begrenses noe av at begge har svært mange år i yrket og også variert erfaring som angår både forskningserfaring og klinisk arbeid. I den grad reliabiliteten mellom deres konklusjoner generaliseres, må dette være til en legespesialistbefolkning med tilsvarende erfaring, som vil utgjøre en undergruppe av legebefolkningen ved OUS.

5.3 Datainnsamling

Datainnsamlingen ble gjort i perioden 2015- 2017 av to uavhenige legegranskere ved OUS, i form av retrospektiv journalgjennomgang med det anerkjente skjemaverktøyet PRISM2. PRISM 2 ble oversatt til norsk etter godkjente kriterier og overført som elektronisk skjema i forbedringssystemet Achilles. uavhengighet ble sikret ved at disse ble utført geografisk og tidsmessig uavhengig og det ikke var kommunikasjon mellom granskerne gjennom skåringsprosessen. Dataene ble overført fra

[51] Ikke for vurderingene

Achilles og gjort tilgjengelig som rådata i Excel der det ble gjort et variabelutvalg som ble overført til Tjenester for sensitive data (TSD) ved Universitetet i Oslo.

5.3.1 Variabelutvalg

Variabelutvalget er hentet fra PRISM 2 (vedlegg 3). Inklusjonskriterene for utvalget var at variabelen var dikotom eller dikotomiserbar og beskrev forebyggbart dødsfall, uønsket hendelse, kvalitet på helsehjelpen eller dokumentasjonsgrunnlaget. Variablene *Alder og Kjønn* ble inkludert for deskriptive metoder og datapresentasjon. Disse er høstet fra administrasjonsystemene, og er lik for begge besvearelser. Variabelen *tidsbruk i minutter* undersøkes for mean tidsbruk per journal for granskerne.

Tabell 11 Variabelutvalg med omforming

	Variabel	Kategorier	Omgjort
1	Alder ved død (år)	Heltall	WHO 5 års aldersintervaller WHO aldersgrupperinger: 0-14 Barn 15-24 Ung voksen 25-64 Voksen 65-74 Ung eldre 75-84 Middels gammel 85+ Gammel
2	Kjønn M/K	Kvinne Mann	
13	Tatt i betraktning alt du vet om denne pasientens innleggelse: Var det noen problemer med helsehjelpen ?	Intet belegg for noe helsehjelpsproblem ⇨ vennligst gå rett til Del D Et visst belegg for helsehjelpsproblem(er) ⇨ svar på neste spørsmål	
14	Finnes det noe belegg for at pasientens død kunne ha vært unngått dersom helsehjelpsproblemet/- ene ikke hadde forekommet?	Nei, dødsfallet var definitivt ikke mulig å unngå ⇨ gå rett til Del D I det minste et snev av belegg for at dødsfallet kunne ha vært mulig å unngå ⇨ fyll ut Del C og så Del D	Variabelen brukes kun til Imputasjon i variabel 16
16	I lys av de problemene i helsehjelpen du har beskrevet over hvor sterkt belegg det er for at dødsfallet kunne ha vært unngått?	Snev av belegg for at det kunne ha vært unngått Mulig unngåelig, men ikke veldig sannsynlig; mindre enn 50-50, men i nærheten Sannsynligvis unngåelig, mer enn 50-50, men i nærheten Sterkt belegg for at det kunne ha vært unngått Kunne definitivt ha vært unngått	Dikotomisert: *Variabel 1* Mer enn 50% sannsynlig Mindre enn 50% sannsynlig
21	Hvordan vil du klassifisere den helhetlige kvaliteten på helsehjelpen? .	Utmerket God Tilstrekkelig Dårlig Svært dårlig	Dikotomisert: Tilstrekkelig eller bedre Utilstrekkelig eller dårligere
22	Ble pasienten utsatt for noen inngripende og uhensiktsmessige prosedyrer ved livets slutt ?	Ja Nei Ikke i stand til å trekke slutning	Parvis sletting av usikkerhetskategori
29	I hvilken grad ga journalen tilstrekkelig informasjon til å bedømme helsehjelpsproblemer?	Pasientjournalene var tilstrekkelige til at man kunne foreta en rimelig bedømmelse Noen mangler Betydelige mangler Alvorlige mangler, umulig å bedømme mulige helsehjelpsproblemer	Dikotomisert: Eventuelle mangler, men tilstrekkelig informasjon til å bedømme Alvorlige mangler, umulig å bedømme
30	Total tid brukt på gjennomgangen (minutter)?	Heltall	

5.3.3 Omforming av variabler

For variabel 22 ble usikkerhetskategorien *Ikke i stand til å trekke slutning* slettet. Det var ingen enighet for denne kategorien, og til tross for at dette er en verdi som mangler ikke- tilfeldig (MNAR), vurderes det at uenigheten som vil beregnes for disse kategoriene er urimelig. Fra Zhao, Liu og Deng (2013) vet vi også at IRR- beregningen for Gwets AC kan øke uten bedret enighet for tomme kategorier, og en kan mistenke at dette også vil gjelde kategorier som brukes i liten grad, ettersom Gwets AC reintroduserer kategorier som en effekt i tilfeldig enighet.

Variablen 16 *Belegg for forebyggbart dødsfall* kollapses til en dikotom variabel med bevisstyrke henholdsvis *over og under 50% sannsynlighet*. Dette kan forsvares med at skillet 50/50 er vanlig i forskningen og at for eksempel skillet mellom *sterkt belegg* og *definitivt unngåelig* kan gi unødig uenighet ved at *definitivt unngåelig* er en svært høy terskel, som heller ikke er hensiktsmessig om målet er implementering av forebyggende tiltak i pasientsikkerhetsarbeidet.

Det bemerkes for den originale variabelen at kategoribeskrivelsen *"Mulig unngåelig, men ikke veldig sannsynlig, mindre enn 50/50, men i nærheten"* oppleves å beskrive noe motstridende konsepter, ved at en sannsynlighet på for eksempel 49 % ikke nødvendigvis vil samstemme med *«ikke veldig sannsynlig»*. Kategoriene ser heller ut til å ikke være fullstendig uttømmende, ettersom kategoriene kun dekker sannsynligheter ≤ .49 og ≥ .51 og granskerne ikke tillates å konkludere med at unngåelig dødsfall er like sannsynlig som usannsynlig (50/50). Det også er en viss avstand mellom de to tilstøtende kategoriene ettersom *snev av belegg og i nærheten* av *50/50*. Dette innebærer at kategoriene ikke nødvendigvis er uttømmende og også at «under 50/50»- variabelen ikke oppleves konsekvent beskrevet, kan gå utover variabelens semantiske validitet og dermed interrater-reliabilitet. For vår omgjorte variabel med over/under 50% sannsynlighet, går vi ut ifra at granskerne først og fremst har forholdt seg til 50/50 skillet.

Rutowski et al. (2019) beskriver at å kollapse ordinale variabler gir informasjonstap og påvirker reliabilitet og presisjon og gir kunstige forbedringer som gjør at konstruktvaliditeten og generaliserbarheten av resultatene kan forringes. En konsekvens av å kollapse variabel 16 *Unngåelig dødsfall* har ført til enighet for to journaler der granskerne i den opprinnelige variabelen var uenige og skåret henholdsvis *Sterkt belegg og Definitivt unngåelig* (Tabell 12, skraverte celler). Ettersom forekomsten av unngåelig dødsfall bestemmes av enighet mellom granskerne, dobles forekomsten for den dikotomiserte variabelen sammenliknet med den ordinale variabelen, fra n=2 til n=4 unngåelige dødsfall. Enighetene har også doblet seg fra n=1 til n= 2 for de to kateogoriene som slås samme til *<50% sannsynlig*, men disse negative funnene har liten innvirkning på

fordelingen etter hot- deck imputeringen og ingen innvirkning på forekomsten. Det kan ikke utelukkes at dette også har forekommet for variablene som angår dokumentasjonskvalitet og helsehjelpskvalitet ettersom den ordinale variabelen kun undersøkes for hovedkonklusjonene

Tabell 12 Endring i enighet og foredling som følge av dikotomisering

	1	2	3	4	5	Total
1 *Snev av belegg for unngåelighet*	0	1	0	0	1	2
2 *Mulig unngåelig*	0	1	0	2	0	3
3 *Sannsynlig Unngåelig*	1	1	**0**	**0**	**0**	2
4 *Sterkt Belegg for unngåelighet*	0	0	**0**	**1**	**2**	3
5 *Definitivt unngåelig*	0	0	**0**	**0**	**1**	1
Total	1	3	0	3	4	11

	<50% sannsynlig	*>50% sannsynlig*	Total
<50% sannsynlig	165	15	180
>50% sannsynig	5	**4**	11
Total	172	19	191

Det har blitt vurdert å dikotomisere variabel 16 også til en dikotom «definitivt / lavere bevisstyrke enn definitivt», men en vet at denne variabelen vil ha lavere IRR og forekomst (n=1) enn 50/50 variabelen. og også at det vanligste for vårt utvalg av forskningslitteratur er 50/50 skillet. Det er usikkert hvor relevant en større bevisbyrde har i pasientsikkerhetsarbeidet, ettersom 50/50 innebærer sannsynlighetsovervekt, noe en vil tro være en robust trigger for å iversette forebyggende tiltak, og granskerne også må oppnå rimelig grad av enighet for at gjennomgangene er nyttige. En unngår stiller spørsmål ved om skillet som ville settes mellom *sterk sannsynlighet for og definitivt forebyggbart dødsfall*[52] er et relevant skille og går ut ifra at det vil være små marginer mellom disse konseptene for mange granskere.

Variabel 21 omhandler generell kvalitet på helsehjelpen, og er dikotomisert till èn variabel med kategorier *tilstrekkelig eller bedre/Utilstrekkelig eller dårligere*. Her er det tatt en frihet i å beskrive kategorien *dårlig* som *utilstrekkelig*, ettersom den tilstøtende kategorien er tilstrekkelig. Det vurderes at det vesentlige er å avgjøre om helsehjelpen er av tilstrekkelig kvalitet, ettersom dette vil dekke forsvarlighetskravet i HPL§4 og Spesialisthelsetjenestelovens §2-2

Variabel 29 Omhandler mangler ved journaldokumentasjonen og er dikotomisert til èn variabel der

[52] Eventuelt også for et skille mellom «nær 50% sannsynlighet»/» « sterk sannsynlighet» vil argumentene være de samme.

kategoriene *Eventuelle mangler, men tilstrekkelig til å bedømme/ Alvorlige mangler, umulig å bedømme, og* kategoriene *noen mangler og betydelige mangler* innlemmes i *«tilstrekkelig til å bedømme»*. Her ansees de første tre kategoriene i den opprinnelige variabelen å ikke være gjensidig utelukkende[53], og et annet skille vil være problematisk for videre analyser. En vurderer her at for vårt reliabilitetsforsøk er det avgjørende om eventuelle mangler i dokumentasjonen gjorde at helsehjelpsproblemene ikke kunne bedømmes.

5.3.4 Manglende data

Manglende kategorisering fra en rater utgjør, ifølge Gwet (2019) og Popper, en ubalansert besvarelse der enighet ikke er mulig, og inklusjon av disse vil underberegne prosent enighet(8, 41). Ettersom sjansekorrigerte reliabilitetskoeffisienter er avhengige av marginale sannsynligheter, som kan endres ved parvis sletting, følges Gwets anbefalning om parvis sletting ved tilfeldig manglende skåring hos èn eller begge granskere.

For variabel *Unngåelighet og Belegg for unngåelighet,* er det et vesentlig antall ikke- tilfeldig (MNAR) og *strukturelt manglende* (SMD) data, Disse følger av at skjemaet gjennomgås kronologisk og skåringsinstruksen rettleder granskeren til å ikke skåre et subjekt i for variabel 14 dersom det er identifisert uønsket hendelse, og dette er tilfelle for gransker A i n=125 journaler, og for gransker B i n= 113 journaler. Videre skal granskerne ikke skåre variabel 16 dersom det ikke er belegg for forebyggbart dødsfall i variabel 14, noe som er tilfelle for gransker A i n=132 og for gransker B i =126 av journalene. Alle funn om mulig unngåelig dødsfall for variabel 14 ble skåret for variabel 16 (n=11).

Instruksen er en logisk følge av at forebyggbart dødsfall ikke kan inntreffe uten en eller flere uønskede hendelser, og at gradering av bevisgrunnlaget for unngåelig dødsfall heller ikke skal skåres dersom det ikke konkluderes med at det er et snev av belegg for forebyggbart dødsfall Mønsteret mellom variablene kan beskrives som *monotont*, etter Andridges (2010) definisjon, ettersom den ene variabelens verdi indikerer den følgende variabelens verdi(55)

En mye brukt tilnærming for strukturelt manglende data i longitudinelle studier, er ifølge Rubin å slette strukturelt manglende data(56). Denne metoden ville se bort ifra at vår skjemainstruks ikke bare forklarer *at* skåringer mangler for forebyggbart dødsfall, men også *hva* verdien ville vært der uønskede hendelser ikke er påvist.

[53] Streng tatt heller ikke de tre siste

En alternativ metode er en mye brukt *deterministisk Hot- deck* imputasjon som kalles *siste observasjon føres videre*[54]. Denne teknikken innebærer, ifølge Andridge (2010), at den manglende verdien erstattes med verdien fra det forrige måletidspunktet ved gjentatte målinger av den samme variabelen(55) Dette utgår fra at den «beste gjetningen» er at verdien ikke har forandret seg siden sist måling(55). Metoden er mye brukt, men som hovedregel ikke lengre anbefalt for de fleste tilfeller ettersom den øker målefeil og kan underestimere effekt av en intervensjon.

En sletting av de strukturelt manglende verdiene hos oss vil ikke påvirke forekomsten av helsehjelpsproblemer, men ville påvirket hver granskermarginalene og en ville reduseret antall skåringer drastisk. Her står en igjen med er en ny variabel som representerer *forebyggbart dødsfall forutsatt at det er påvist en uønsket hendelse*. Denne variabelen kan forventes å ha en jevnere fordeling, men en fjerner samtidig de fleste enighetene, og en kan også få større variasjon ettersom utvalget reduseres, eventuelt også ikke- signifikante resultater. Lav IRR er antagelig også rimelig for en slik variabel, ettersom den forutsetter enighet for helsehjelpsproblemer, og disse subjektene bør være enklere å skåre slik at uenighet også skal «straffes» mer. Et annet hensyn som er avgjørende for vurderingene for oppgaven er at den betingede variabelen beskriver en undergruppe og ikke er like nyttig som den opprinnelige når det gjelder beregning av forekomsten av forebyggbare dødsfall i utvalget.

For oppgaven vurderes det at *siste observasjon føres videre*- tilnærmingen kan benyttes for utregning av IRR, ettersom en ikke finner litteratur som beskriver håndtering av vår spesielle type manglende data, gjøres også en beregning med 50/50 sannsynlighet, der alle manglende verdier slettes parvis. Tilfeldig manglende verdier fra variabel 13 og 14, som Andridge (2010) ville omtalt som *donorvariablene*(55), videreføres ellers som manglende verdier i de andre to variablene, og slettes parvis for IRR- analysene. Der parvis sletting kan påvirke fordelingen uhensiktsmessig, bemerkes dette i resultatdelen for hver variabel.

For variabelen *unngåelig dødsfall* hentes andelen strukturelt manglende verdier fra *helsehjelpsproblemer*- variabelen, og der det er funnet helsehjelpsproblemer, ser en på variabel 14, *belegg for unngåelig dødsfall*. Dersom det for denne er konkludert med definitivt ikke unngåelig dødsfall, imputeres dette for unngåelig dødsfall – variabelen. Tilfeldig manglende verdier som mangler helsehjelpsproblemer videreføres som tilfeldig manglende. Dersom det mangler verdier i helsehjelpsproblemer men disse er skåret i belegg for unngåelig dødsfall, overføres skåringen.

[54] *Last observation carried forward (LOCF). Hot- deck* er betegnelsen for metoder der en benytter en donor- variabel for imputasjon der variablene er liknende eller overlappende (40)

Variabel 14 er for oss overflødig ettersom verdiene imputeres i den nye dikotomiserte variabel 16 og det utføres ikke analyse av IRR for denne.

5.4 Statistiske metoder og presentasjonsvalg

Først presenteres utvalgsfordelingen med kjønnsfordeling med sentralitetsmål samt WHO alderskategoriseringer med tabeller. Deretter sammenfattes resultatene fra hovedanalysene i en tabell med IRR- beregning per variabel. Fokus for analysen er beregning av interrater- reliabilitet for variablene *Problemer i helsehjelpen og Unngåelig dødsfall.* Prosent enighet samt de tilfeldighetskorrigerte IRR- koeffisientene Cohens Kappa i IBM SPSS og Gwets AC1 i R Studio med irrCAC pakken.

Variablene *Tidsbruk, Inngripende* og *Uhensiktsmessige prosedyrer* samt *Dokumentasjonskvalitet* ble inkludert for analyser av hensyn til kontekstualisering av hovedfunnene for diskusjonskapitlet. IRR- beregningene for de to kategoriske variablene oppgis i resultatdelen, ettersom disse ga beregninger for Cohens Kappa som er spesielt interessante for vurderingen av Kappa ved skjeve fordelinger. Ellers beskrives de mest relevante resultatene fra dissse analysene kort i diskusjonsdelen der disse er relevante for hovedfunnene, og analysene er ellers vedlagt (vedlegg 6) ettersom variablene ikke er direkte relevante for oppgavens forskningsspørsmål, og teorikapitlet ikke er skrevet med tanke på å forstå skåringen av disse.

Signifikansnivå settes til 95% for alle analyser. SPSS angir standardfeil (SE), og konfidensintervall beregnes fra SE etter IRR -/+ (1.96 x SE). R Studio oppgir konfidensintervall direkte. P – verdi oppgis som <.05 eller <.001 der IRR er signifikant, og P- verdien skrives ut der IRR ikke er signifikant. Enighet og reliabilitet presenteres sammenfattet i 2x2 enighetstabeller per kjønn og for utvalget som helhet, med en repetisjon av IRR- målene med konfidensintervall og p – verdi.

Tendenser for granskernes tidsbruk per journal undersøkes for hvert kjønn, utvalget som helhet og hovedkonklusjonene Students t- test og point biserial korrelasjon etter at variabelene er gjort normalfordelte (Log 10).

For de originale ordinale variablene dokumentasjonskvalitet og belegg for forebyggbart dødsfall, utføres Wilcoxon signed rank test for å undersøke forskjeller mellom granskerne i observerte skår. Z- score for denne testen oppgis ikke, ettersom denne vil tolkes som effektestimat mellom to måletidspunkt og vil være misvisende som variasjonestimat for enighet mellom forskjellige granskere.

For de to dikotomiserte variablene, beskrives den opprinnelige ordinale kun med enighetstabeller og og forskjeller i skår testes med Wilcoxon rank order. Effektestimater oppgis ikke for denne,

ettersom testen er best egnet for to måletidspunkt og z- score vil være misvisende som variasjonestimat for enighet mellom forskjellige granskere.

For aldergruppene sammefattes funnene i en tabell med alle variabler og enigheter/ uenigheter. Koeffisienter ble forsøksvis regnet ut også for WHO aldersgruppene, men flere av gruppene har svært få observasjoner og ettersom flere av IRR- estimatene ble ikke- signifikante ble det konkludert med at det er uhensiktsmessig å sammenlikne disse.

Ekskluderte koeffisienter

Koeffisenter som ikke kan kalkuleres med allment tilgjengelig statistikkprogrammer utelukkes i første omgang. Det finnes svært mange koeffisenter og undervarianter å velge mellom, og Gwet (2008) besrkiver at litteraturen drukner i et mangfold av prosedyrer uten noe felles rammeverk for å evaluere deres validitet(41) og Kottner (2011) og Zhao (2013) påpeker at manglende tydeliggjøring av målenes matematiske og konseptuelle grunnlag gjør det vanskelig å begrunne koeffisientvalget (46, 47).Det begrunnes derfor kort hvorfor relevante og populære koeffisienter egnet for vårt design med dikotomt målenivå og to granskere ekskluderes, og det kan ikke garanteres at alle relevante koeffisienter er vurdert.

Scotts Pi ekskluderes ettersom den ifølge Gwet (2012) er svært sensitiv for forekomsten av trekk ettersom fordeling er en hovedfaktor i beregning av tilfeldig enighet for koeffisienten. Dette er ifølge Gwet en svakhet da indeksen skal måle enighet men ikke forekomst(41). Den er også basert på at alle kategorier har lik sannsylighet for å brukes, noe som ifølge Zhao, Liu og Deng (2013) er en urimelig antagelse i forskning

Fleiss`Kappa ekskluderes ettersom den ifølge Hallgren (2012)er upassende i fullt kryssede design ettersom den utgår fra en antagelse om at hvert subjekt får et nytt tilfeldig utvalg granskere(27)

Krippendorffs alpha ekskluderes ettersom den ikke anses å gi noe korrektiv til Cohens Kappa. Hos Zhao, Liu og Deng (2013) beskrives Krippendorffs alpha for vårt design og antatte fordeling som like konservativ som Cohens Kappa og vil produsere urimelig lav IRR etter den samme mekanismen som Kappa (28, Tabell 19.12 og 19.13)

Prevance Adjusted, Bias Adjustet Kappa (PABAK) også kjent som *Brennan Predigers koeffisient* ble ikke vurdert, ettersom den ikke er tilgjengelig i SPSS og ikke så ut til å være tilgjengelig i R –

studio når koeffisientvalget ble tatt. Den er senere funnet å være tilgjengelig i pakken epiR og kan vurderes som alternativ til kappa for tilsvarende oppgaver.

6.0 Veiledere og ressurser

Masteroppgaven inngår i et forskningsprosjekt ved Avdeling for Pasientsikkerhet og Kontinuerlig Forbedring ved OUS. Biveileder er Dr. med. Trine Sand Kaastad som leder avdelingen og er prosjektleder for studien. Hovedveileder er Dr. med Anne Karin Lindahl, avdelingsdirektør for avdeling Kvalitet og Pasientsikkerhet ved Kunnskapssenteret i Folkehelseinstituttet, avdelingsdirektør kirurgisk avdeling ved AHUS og førsteamanuensis ved avdeling for helseledelse og helseøkonomi ved UiO. Statistikkprogrammene SPSS og R Studio samt Micorsoft EXCEL og WORD med referanseprogrammet ENDNOTE benyttes i Tjenester for Sensitive data (TSD) gjennom tilgang som Student ved UIO.

6.2 Etiske hensyn og personvern

Hovedstudien er en kvalitetsstudie og unntatt vurdering i REK. Dispensasjon fra samtykke og godkjenning ble gitt av Personvernombud OUS i 2015 (Vedlegg 2). Masteroppgaven går i sin helhet under delmål for hovedstudien. Norsk senter for forskningsdata (NSD) ble kontaktet på telefon våren 2021 og det ble avklart at oppgaven ikke var meldepliktig der. Datatilgang for studenten er godkjent i personvernombud 19/9 2021(Vedlegg 1), og databehandling foregår via innlogging som ansatt i OUS på server / K-Område, via privat wifi med bank- ID innlogging. og på TSD tjenester for sensitive data ved UIO med innlogging ved Google Authentication samt brukernavn og passord.

Rådata er avidentifisert med ID - nummer tilknyttet hver journal. I datasettene inngår personopplysninger koblet til sensitive helseopplysninger, og ID- numrene kan tilbakeføres til person via en fysisk adskilt kodeliste. Oppgaveskriver har ikke tilgang til kodelisten og oppgir at all omgang med dataene har vært i tråd med aktsomhetskravet i §4 i Forskningsetikkloven(57). Studenten jobber som spesialsykepleier i OUS, i et arbeidsforhold uten tilknytning til hovedstudien og oppgir ingen interessekonflikter. Oppgaven mottar ikke finansiering. Det er kjent at biveileder Trine Sand Kaastad var en av legegranskerne som utførte journalgjennomgangen.

6.0 Resultater

I dette kapitlet presenteres først fordeling over alder ved WHOS aldersklassifiseringer samt kjønnsfordeling for utvalget. Deretter følger hovedkonklusjonene for utvalget samt for menn og kvinner separat, og til slutt legges analysene for det resterende variabelutvalget fram. For alle kvalitative tolkninger av IRR benyttes Landis og Kochs (1991) intervaller. Prosenter beregnes etter

totalt antall journaler i gruppen, og manglende verdier oppgis separat. Prosent enighet fra IRR-tabellen er oppgitt relativt til de journaler som ble skåret for variabelen.

6.1 *Alder* og *Kjønn* i utvalget

Den eldste personen i vårt utvalg er en kvinne på 97 år, men aldersgjennomsnittet for kvinnene i utvalget ligger noe under aldersgjennomsnittene for utvalget (Mean 65.9 år, SD 23.6) mens menn ligger noe under. Det er færre kvinner enn menn i utvalget (n=93 (46.3%) / n=107 (53.5%)) Median alder i utvalget er 72 år, og ligger over aldersgjennomsnittet i alle grupper, men for menn er forskjellen noe mindre enn for kvinner. Alle fordelinger på WHOs aldersgrupper er bimodale[55], og både kvinner og menn har opphopninger i Aldersgruppen «Gammel», som er en gruppe som strekker seg over 12 år til maksimal alder i utvalget. Det kan se ut som om fordelingene ville vært unimodal dersom gruppen «middels gammel» (75-84år) ikke var underrepresentert. Kvinner har en opphopning i aldersgruppen ung voksen (15-24år). Det er som forventet få barn (0-14 år) i utvalget. I de to yngste og de to eldste aldersgruppene har kvinner høyere representasjon, og i de to midtre er det flere menn.

Tabell 13 WHO aldersgrupper og sentralitetsmål

		Alle		Kvinner		Menn	
Alder	**Aldersgruppe**	n	%	n	%	n	%
0 - 14	Barn	12	6 %	6	6.4%	6	5.6%
15-24	Ung voksen	43	21.5%	22	23.6%	21	19.6%
25-64	Voksen	44	22 %	19	20.3%	25	23.3%
65-74	Ung eldre	47	23.5%	19	20.3%	28	26.1%
75-84	Middels gammel	5	2.5%	3	3.2%	2	1.8%
85+	Gammel	49	24.5%	24	25.8%	25	23.3%
	Total	200	100 %	93	100 %	107	100 %

	Alle	**Kvinner**	**Menn**
n	200	93	107
Min	0	0	0
Max	97	97	96
Median	72	72	71
Typetall 1 i gruppe	Gammel	Gammel	Ung eldre
Typetall 2 i gruppe	Ung eldre	Ung voksen	Gammel
Mean	65.9	64.5	67.1
SD	23.6	25	22.4

6.2 Oppsummering Interrater- reliabilitet

Gwets AC1 Ligger nær opp mot prosent enighet og innen den kvalitative tolkningen Almost perfect for alle variabler, unntatt for variabelt 13 *Problem i helsehjelpen*, der AC1 er en del lavere og innen den samme kvalitative tolkningen som Cohens Kappa (*Fair).* Ingen av beregningene av Kappa når

[55] WHOS aldersgrupper reflekterer livsfaser og er ikke- linjære intervaller. Beskrivelser av representasjon i gruppene er avlest og beskrevet utifra tabellene, og det er ikke undersøkt om over/ underrepresentasjon er statistisk signifikant.

opp til moderat reliabilitet, og unntatt for *Problem i helsehjelpen*, er nedre konfidensintervall for Kappa for alle variabler kappa under 0, som tilsvarer at en må konkludere med tolkningen *Poor,* eller også dårligere reliabilitet enn ved tilfeldig gjetning dersom en regner inn feilestimatet. For variabel 22 *Inngripende eller uhensiktsmessige* prosedyrer, er Kappa ikke signifikant. Dette tolkes som at Kappa for utvalgsstørrelsen ikke er signifikant høyere enn IRR = 0 / tilfeldig enighet. For variabelen 29 *Dokumentasjonskvalitet* kunne ikke Kappa beregnes ettersom Gransker A vurderte at alle variabler hadde tilstrekkelig dokumentasjonskvalitet. Samtidig er Prosent enighet og Gwets AC1 er like (98.4% og .984) svært nær perfekt enighet for denne variabelen.

Tabell 14 Prosent enighet, Cohens Kappa og Gwets AC1 for hovedkonklusjonene

			Cohens			Gwets				Tolkning	
Nr.	Variabelnavn	% Enighet	Kappa	KI	Sig.	AC1	KI		Sig.	Kappa	Gwets AC
13	Problem i helsehjelpen	65.4%	.267	.128	.406	<.001	.384	.209	.488	<.05	Fair	Fair
16	Unngåelig dødsfall*	88.5%	.209	-.006	.424	<.001	.865	.806	.925	<.001	Fair	Almost perfect
21	Kvalitet av helsehjelpen	88.9%	.118	-.066	.302	<.05	.874	.819	.929	<.001	Slight	Almost perfect
22	Inngripende (…) prosedyre	86.7%	.026	- .111	.163	.664	.846	.783	.910	<.001	Slight	Almost perfect
29	Dokumentasjonskvalitet	98.4%	**				.984	.965	1	<.000	**	Almost perfect

*Hot-deck imputasjon

** Kappa kunne ikke beregnes

6.3 Helsehjelpsproblemer

Gransker A fant et visst belegg for helsehjelpsproblemer i 33% av journalene (n= 66) og Gransker B fant det samme i 39% av Journalene (n=78). Det var enighet mellom granskerne om kategorien at det var et visst belegg for helsehjelpsproblemer i 19.5% av journalene (n=39), og enighet om at det ikke var belegg i 43% av journalene (n =86). Granskerne var uenige i 33% av journalene (n=66)

Observert enighet (Tabell 15) var 65.4%, Kappa var .268, som tolkes kvalitativt som *fair,* mens nedre konfidensintervall ligger innen kategorien *slight* (.128). Gwets AC1 var .384 som tolkes som *fair,* med *n*edre konfidensintervall innen den samme kategorien (.209).

Manglende verdier n=9, 4.5%

Tabell 15 Helsehjelpsproblemer

		Gransker B		
		Intet belegg	*Et visst belegg*	Total
Gransker A	*Intet belegg*	86	39	125
	Et visst belegg	27	39	66
	Total	113	78	191

6.4 Unngåelige dødsfall- ordinal variabel

Wilcoxon signed rank order viste at gransker B skåret høyere sannsynlighet for unngåelig dødsfall enn gransker A i 34 journaler og lavere enn gransker A i 8 journaler. Det var enighet mellom granskerne i 146 journaler. Forskjellene var signifikante (*P*- verdi .002).

Prosent enighet (observert enighet/ *n = 3/11*) var 27.3% Det var enighet mellom granskerne for èn journal per kategori for de tre kategoriene *sterkt belegg*, *sannsynlig unngåelig* og *mulig unngåelig.* Det var ingen enigheter for kategorien *Sannsynligvis unngåelig*, og gransker B skåret ingen journaler i denne kategorien. Det var heller ikke noen enighet for kategorien *snev av belegg*.

Figur 4 Enighetstabell unngåelig dødsfall (ordinal)

		Gransker B					
		1	**2**	**3**	**4**	**5**	**Total**
Gransker A	**1** *Snev av belegg for unngåelighet*	0	1	0	0	1	2
	2 *Mulig unngåelig*	0	1	0	2	0	3
	3 *Sannsynlig Unngåelig*	1	1	0	0	0	2
	4 *Sterkt Belegg for unngåelighet*	0	0	0	1	2	3
	5 *Definitivt unngåelig*	0	0	0	0	1	1
	Total	1	3	0	3	4	11

*Alle som ble skåret *snev av belegg* i variabel 14 er skåret (n = 11).

6.5 Unngåelig dødsfall- dikotomisert

Gransker A fant over 50% belegg for unngåelig dødsfall i 5.5% Journalene (n= 11) og Gransker B fant belegg for unngåelig dødsfall i 9.5% av Journalene (n=19). Det var enighet mellom granskerne om kategorien at det hadde forekommet unngåelig dødsfall i 2% av journalene (n=4), og enighet om at dødsfallet ikke var unngåelig hos 82.5% av journalene (n =165). Granskerne var uenige i 10% av journalene (n=20)

Observert enighet (Tabell 12) for unngåelig dødsfall var 88.5%, Kappa var .209, som tolkes kvalitativt som *fair*, med nedre konfidensintervall innen tolkningen *slight* (.006). Gwets AC1 var .865 som tolkes som *nesten perfekt*, med nedre konfidensintervall innen tolkningen *Substantial* (.806).

Tabell 16 Unngåelig dødsfall

Mangler n=9, 4.5%

		Gransker B		
		<50% sannsynlig	*>50% sannsynlig*	Total
Gransker A	*<50% sannsynlig*	165	15	180
	>50% sannsynig	5	4	11
	Total	172	19	191

6.6 Hovedkonklusjoner Menn

Gransker A fant et visst belegg for helsehjelpsproblem hos 30%% (n=32 av menns journaler, og granskers B fant det samme hos 47% (n =48). Det var enighet mellom granskerne om at det var et visst belegg for helsehjelpsproblem i 23.5% (n=24) av journalene, og enighet om at det ikke var et slikt belegg i 45.1% (n=46) av journalene.

Observert enighet for helsehjepsproblemer innen gruppen menn var 70.6%, Kappa var .358, som tolkes kvalitativt som *fair*, med nedre konfidensintervall innen samme tolkning(K=-244). Gwets AC1 var .400 som tolkes som *fair* og grensende til *moderate*, med nedre konfidensintervall innen tolkningen *fair* (.215).

Gransker A fant at det var over 50% sannsynlighet for unngåelig dødsfall hos 4.9% av menns journaler(n=5), og gransker B identifiserte slike hendelser hos 11.7 % (n=12). Det var enighet mellom granskerne om at det var over 50% sannsynlig at et unngåelig dødsfall hadde forekommet i 2.9% av joruanlene (n=3), og enighet om at det var under 50% sannsynlig at et unngåelig dødsfall hadde forekommet hos 86.2% (n=86.3%). Granskerne var uenige i 10.8% av journalene (n=11)

Observert enighet for unngåelig dødsfall innen gruppen menn var 89.2%, Kappa var .305, som tolkes kvalitativt som *fair*, med nedre konfidensintervall innen tolkningen *sligh*t (.009). Gwets AC1 var .872 som tolkes som *Almost perfect*, med nedre konfidensintervall innen tolkningen *Substantial* (.793).

Manglende verdier n=5 (4.7)

		Gransker B		
		Intet belegg	*Et visst belegg*	Total
Gransker A	*Intet belegg*	46	24	70
	Et visst belegg	8	24	32
	Total	54	48	102

		Gransker B		
		<50% sannsynlig	*>50% sannsynig*	Total
Gransker A	*<50% sannsynlig*	88	9	97
	>50% sannsynig	2	3	5
	Total	90	12	102

Nr.	Variabel	% Enighet	Kappa	KI		Sig.	AC1	KI		Sig.
13	*Problem i helsehjelpen*	70.6%	.358	.244	.471	<.001	.400	.215	.586	<.001
16	*Unngåelig dødsfall*	89.2%	.305	.009	.600	<.001	.872	.793	.952	<.001

6.7 Hovedkonklusjoner Kvinner

Gransker A fant et visst belegg for helsehjelpsproblemer i 36.5% i kvinners journaler (n= 34) og Gransker B fant det samme i32.2% av journalene (n=30). Det var enighet mellom granskerne om at det var et visst belegg for helsehjelpsproblemer i 16.1% av kvinners journaler (n=15), og enighet om at dødsfallet ikke var unngåelig hos 43% av journalene (n =40). Granskerne var uenige i 36.5% av journalene (n=34)

Observert enighet for helsehjepsproblemer innen gruppen kvinner var 61.7%, Kappa var .172, som tolkes kvalitativt som *Slight.* Nedre konfidensintervall tilsier kvalitativ tolkning *Poor* (-.035). Gwets AC1 var .291, som tolkes som *Fair* og nedre konfidensintervall tilsier tolkningen *Slight* (.078)

Gransker A fant over 50% belegg for unngåelig dødsfall i 6.4% kvinners journaler (n= 6) og Gransker B i 7.5% av Journalene (n=7). Det var enighet mellom granskerne om at det hadde forekommet unngåelig dødsfall i 1.1% av journalene (n=1), og enighet om at dødsfallet ikke var unngåelig hos 82.7% av journalene (n =77). Granskerne var uenige i 11.8% av journalene (n=11)

Observert enighet for unngåelig dødsfall innen gruppen kvinner var 77%, Kappa var .088, som tolkes kvalitativt som *slight,* med nedre konfidensintervall innen tolkningen poor (-.194). Gwets AC1 var .857 som tolkes som *nesten perfekt,* med nedre konfidensintervall innen tolkningen *Substantial* (.767).

Tabell 17 Helsehjelpsproblemer Kvinner

Manglende verdier for begge variabler er n=4, 4.3%

		Gransker B		
		Intet belegg	*Et visst belegg*	Total
Gransker A	*Intet belegg*	40	15	55
	Et visst belegg	19	15	34
	Total	59	30	89

Tabell 18 Unngåelig dødsfall Kvinner

		Gransker B		
		<50% sannsynlig	*>50% sannsynig*	Total
Gransker A	*<50% sannsynlig*	77	6	83
	>50% sannsynig	5	1	6
	Total	82	7	89

Tabell 19 IRR hovedkonklusjoner Kvinner

Nr.	Variabel	% Enighet	Kappa	KI		Sig.	AC1	KI		Sig.
13	*Problem i helsehjelpen*	61.7%	.172	-.035	.379	.102	.291	.078	.506	<.05
16	*Unngåelig dødsfall*	77 %	.088	-.192	.368	.407	.857	.767	.947	<.001

Resterende analyser av variabler som ikke angår hovedkonklusjonene direkte *(Generell helsehjelpskvalitet, Dokumentasjonsgrunnlag, Inngripende og uhensiktsmessige prosedyrer samt Tidsbruk) er* vedlagt (vedlegg 6).

7.0 Diskusjon

I diskusjonskapitlet beskrives først aspekter ved vårt utvalg før resultatene fra analysene og protokollen for vår journalgjennomgang diskuteres med bakgrunn i teoriene presentert i oppgavens første del. Forskningsspørsmål 1 er besvart i resultatkapitlet, men funnene kontekstualiseres i med funn fra tidligere studier.

Forskningsspørsmål:

1. Hva er forekomsten av uønskede hendelser og forebyggbare dødsfall for utvalget, og i hvor mange journaler er det enighet for disse vurderingene?

2. Hvordan kan tendenser for variasjon i skår mellom granskerne forklares?

3. Hva er interrater- reliabilitet målt ved Prosent enighet, Cohens Kappa og Gwets AC1 for variabelutvalget

4. Kan forskjeller mellom målene forklares av deres følsomhet for datafordelingen?

Median alder for vårt utvalg var 72 år, og selv om en kan anta at medianalder for alle inneliggende pasienter på sykehus vil ligge et stykke over median alder i Norge, som var 39 år i 2014 (58), kan en anta at pasienter som døde i sykehus vil være en eldre undergruppe av sykehusbefolkningen. Medianverdi ligger over gjennomsnitt i alle grupper noe som tyder på at utvalgsfordelingen er venstreforskjøvet,noe som innebærer at de fleste i utvalget dør i en høy alder. Ved undersøkelse av boxplot sees også at de som dør i alderen 0- 15 år er statistiske ekstremskårere/outliere for utvalget.

Dersom en sammenlikner vårt utvalg med tall fra Kalseth og Halvorsens (2020) undersøkelse av hvor norske borgere døde i årene 2003- 2011 (N≈ 350 000) (59), ser en at utvalget vårt har noe flere menn enn utvalget fra det Norske dødsårsaksregisteret (53,5% vs 47%). Den eldste personen i vårt utvalg er en kvinne på 97 år, men aldersgjennomsnittet for gruppen kvinner ligger noe under gjennomsnittene for utvalget og gruppen menn.

Dette kan forklares av tall fra Kalseth og Halvorsens studie, som viste at kvinner i større grad enn menn døde på sykehjem(59). Sannsynligheten for dødsfall på sykehjem var svært høy for demens(59),og selv om forfatterne ikke undersøkte sammenhengen for kvinner og menn, kan en se for seg at sammenhengen mellom høy alder og demens også påvirker at færre kvinner dør på sykehus. Norske kvinner har høyere forventet levealder enn menn, og dersom en utgår fra at de som når en høyere alder vil ha dårlige prognoser og liten nytte av behandling ved alvorlig sykdom, kan dette bidra til å forklare den lavere representasjonen av kvinner i utvalget

Alle fordelinger på WHOs aldersgrupper er bimodale[56], og både kvinner og menn har opphopninger i Aldersgruppen «Gammel», som er en gruppe som strekker seg over 12 år til maksimal alder i utvalget, eventuelt kan en se det som at gruppen «middels gammel» (75-84år) er underrepresentert. En kan tenke seg at flere av de som lever til en høy alder i utgangspunktet har hatt god helse i alderdommen og oftere legges inn i sykehus fra hjemmet enn å dø under opphold på sykehjem. Kalseth og Halvorsen (2020) fant også at flere i aldersgruppen 67- 79 år døde på sykehjem enn i sykehus(59), noe som samsvarer med en slik antakelse. Imidlertid viste studien også at aldersgruppen 80-89 år oftere døde hjemme enn på sykehus, og det at denne gruppen allikevel er godt representert i vårt utvalg, kan antagelig forklares av at hele 60% av dødsfall rammer personer over 80år(59).

[56] Det bemerkes at utvalgetsfordelingen er vurdert etter WHOS aldersgrupper, som reflekterer livsfaser og er ikke- linjære intervaller. Beskrivelser av representasjon i gruppene er avlest og beskrevet utifra tabellene, og det er ikke undersøkt om over/ underrepresentasjon er statistisk signifikant.

Det sees en opphopning i aldersgruppen ung voksen (15-24år) i vårt utvalg, noe som er påfallende. Kanskje kan dette ha sammenheng med at Regional Seksjon for Spiseforstyrrelser ligger under OUS, eller annen overrepresentasjon av alvorlig sykdom der unge kvinner er overrepresentert på grunn av regionalisering av helsehjelpstilbudet.

Det er som forventet få barn (0-14 år) i dette utvalget, kvinner er lett overrepresentert også i denne gruppen, og ved en nærmere titt på dataene ser en at gruppen barn preges av barn <1 år som utgjør 4.3% innen gruppen kvinner og 2,8% innen gruppen menn. I de neste to aldersgruppene (25 – 74 år) er menn overrepresentert. En kan tenke seg at dette kan ha å gjøre med at unge voksne menn kan være mer utsatt for ulykker enn unge kvinner, og at menn i gruppen *voksen* kan ha høyere risiko for eksempel relatert til hjerte- og karsykdom enn kvinner. I tråd med denne antagelsen, fant Kalseth og Halvorsen at det er få dødsfall på sykehjem relatert til eksterne årsaker eller hjerte – og karsykdom(59). En annen observasjon, er at det kan se ut til at det er 0.3% «farligere» å være mann i alderen 65-74år, enn å være kvinne over 85år. Gjennomsnittlig forventet levealder for menn i 2014 var 80år og for kvinner noe over 84år, og det vil være langt færre menn i aldersgruppen over 85år, samtidig som mange i henhold til forventet levealder vil dø «middels gamle» og dermed oftere i annen helseinstitusjon eller hjemme.

Korrelasjonsanalyser for sammenhengen mellom alder og konklusjon om forebyggbart dødsfall var ikke signifikante for vårt utvalg, det vil si at forekomsten av unngåelig dødsfall var jevnt fordelt på aldersrepresentasjonen i utvalget. Dette antas i ha med utvalget å gjøre, ettersom utvalget var venstreforskjøvet og de fleste unngåelige dødsfallene også inntraff i de eldre og «største» aldersgruppene. Alt i alt ser utvalget vårt til å stemme godt overens med Kalseth og Halvorsens (2020) beskrivelser av hvem som dør i sykehus, noe som tyder på at vi har tilfredstillende *utvalgsvaliditet* og utvalget etter Poppings (2019)definisjon er representativt, og dermed egnet for generalisering av resultatene(44) til den større befolkningen som dør på sykehus. Kalseth og Halvorsen beskriver at over 40% av dødsfallene i deres studieperiode inntraff på sykehus, noe som innebærer at studier med vår utvalgsmetodikk potensielt kan generaliseres til rett under 16 000 årlige sykehusdødsfall[57].

Hogans (2016)beskrivelser av aspekter som kan gi utvalgsskjevhet der inklusjonskriteriet er at pasienten har avgått med døden under sykehusopphold(19), er svært aktuelle også for vår studie. Vi kjenner til assosiasjonen mellom kroniske lidelser og uønskede hendelser fra Amalberti et al.(2011) (17) og også mellom høy alder og helsehjelpsproblemer blant annet hos Brennan et al.(36), og dersom risikoen for unngåelig dødsfall øker med akkumulasjonen av helsehjelpsproblemer, kan

[57] N= 350 000 x 41% / 9 år

sykehusbefolkningen består av distinkte undergrupper med svært ulik risiko også for unngåelige dødsfall.

En begrensning ved vårt utvalg er at vi ikke kan vite om de som er funnet å være rammet av uungåelig dødsfall er personer som uansett ville dødd på sykehus, eller om de er en undergruppe av den desidert største andelen sykehuspasienter, som skrives ut i live der helsehjelpen er forsvarlig.

PRISM2 innhenter riktignok legegranskernes vurdering av hvor mye livet ble forkortet av helsehjelpsproblemene i kategorier fra *èn uke eller mindre* til *Antall år,* ut ifra en skjønnsmessig vurdering av pasientens prognoser forutsatt tilstrekkelig helsehjelp. Ettersom vårt utvalg inkluderer palliative pasienter der dødsfallet var forventet, kunne en antagelig erstattet denne variabelen med en vurdering av om pasienten med tilfredsstillende helsehjelp ville overlevd oppholdet og perioden etter utskrivelse, som hos Hayward og Hofer(2001)(35). Det er ikke innlysende at et anslag av forkortet livslengde bidrar til forståelsen av problemer eller forebygging av disse, ettersom den forventede livslengden for de fleste i et slikt utvalg vil være avhengig av helsetilstand og alder ved innleggelsen.

Det at ingen pasientgrupper er ekskludert i vårt utvalg kan være en styrke eller en svakhet. Vi ser at liknende studier gjerne ekskluderer enten pasientgrupper med svært høy dødelighet, som hos Hayward og Hofer (2001) (35), eller pasientgrupper med svært lav dødelighet som hos Hogan(2014) (16). Det siste gjøres ettersom en ifølge Hogan (2016) måler *forekomsten av sjeldne hendelser innen en svært sjelden hendelse(19).*Det å ikke ekskludere pasientgrupper med svært lav dødelighet kan tenkes å øke forekomsten av det Hogan (2016) beskriver som *støy som overdøver et allerede lavt signal*(19). Vårt skjemainstrument ble utviklet av Hogan et al (2012, 2014-15) for studier der psykiatriske, obstetriske og pediatriske pasienter ble ekskludert (14, 16, 60), men det er ikke beskrevet om en slik utvalgsmetodikk er en forutsetning for skjemaets validitet. Vi vet heller ikke hvilke kriterier våre legegranskere legger til grunn for å vurdere standarden for faglig praksis innen psykiatriske journaler, som kan sies å stå i en noe annen vitenskapelig tradisjon enn den kliniske medisinen. Vi vet heller ikke om journalgjennomgangsmetoden er like godt egnet for deteksjon av helsehjelpsproblemer for alle pasientgrupper.

For vårt utvalg ble det ble funnet et visst belegg for helsehjelpsproblemer i 33%, og over 50% sannsynlighet for unngåelig dødsfall i 2.9% av journalene. En ser at vårt utvalg er en del yngre (65.9 år vs 77år) enn utvalget for Rogne et als studie fra 2019, der det ble funnet en noe høyere forekomst på 4.2% unngåelige dødsfall med utgangspunkt i den samme definisjonen. Kjønnsfordelingen for utvalgene kan beskrives som sammenliknbar (46.5 vs 45%). Det yngre utvalget hos Rogne et al. kan kanskje forklares av at pasienter under 16 år og psykiatriske pasienter som antas å ha lavere alder enn gjennomsnittet for sykehusdødsfall er ekskludert.. Den mest

åpenbare forklaringen vil også her ha sammenheng med alderen i utvalget, og kanskje også at eksklusjonen av psykiatriske pasienter i seg selv ekskluderte en del journaler som en ser for seg kan være vanskelige å skåre. I sammenlikning med andre norske studier ser en at vårt utvalg er en del yngre (65.9 år vs 77år), mens kjønnsfordelingen er relativt lik (46.5 vs 45%) (5). Det yngre utvalget hos Rogne et al kan muligens forklares av at pasienter under 16 år og psykiatriske pasienter som kan antas å ha lavere alder enn gjennomsnittet for sykehusdødsfall er ekskludert. Rogne et al. fant en noe høyere forekomst på 4.2% unngåelige dødsfall mot våre 2.9%, og la den samme definisjonen av over 50% sannsynlighet til grunn som for vår studie. Den mest åpenbare forklaringen vil også her ha sammenheng med alderen i utvalget, og kanskje også at eksklusjonen av psykiatriske pasienter i seg selv ekskluderte en del journaler som en ser for seg kan være vanskelige å skåre.

I Deilkås`studie fra 2015 ble ikke forekomsten av unngåelige dødsfall bestemt, men forekomsten av de alvorligste uønskede hendelsene ble funnet å være 13%. Ettersom datagrunnlaget er en sammenfatning av tall fra forskjellige sykehus og forekomsten for hver av disse ikke er presentert, er det vanskelig å vurdere hvor mye variasjon forekomsten har påvirket dette tallet. En kan heller ikke gjøre et rimelig anslag av hvor mange av de alvorligste hendelsene som førte til unngåelig dødsfall Global Trigger Tool ser ut til å være et langt mindre omfattende instrument enn PRISM 2 etter beskrivelser fra Hibbert et al (2016) (26)

For Harvard Medical Practice Study (30) var forekomsten av sannsynlig unngåelige dødsfall 6%, og om en utgår fra at «sannsynlig» tilsvarer vår definisjon ved 50% sannsynlighetsovervekt, kan forklaringen på denne relativt høye forekomsten være at resultatet av utvalgsmetodikken i denne studien var at over halvparten av journalene i utgangspunktet hadde høy risiko for helsehjelpsproblemer. Terminalt syke var imidlertid ekskludert fra utvalget, men ettersom de utgjorde en liten andel av befolkningen (n= 66) og snittalderen til tross for dette er noe høyere enn i vårt utvalg er det usikkert om dette forklarer forskjeller mellom studiene. Det er heller ikke gitt at hvordan det at en pasient i utgangspunktet er terminalt syk vil påvirke legers vurderinger av unngåelig dødsfall, Reliabilitetsberegningen med ICC .24-.34 ligger over Kappa for vår tilsvarende variabel (.209), men et godt stykke under AC1 (.860). For Kappasammenlikningen kan en tenke seg at utvalgsstørrelsen hos Hayward og Hofer i seg selv har gjort reliabiliteten bedre enn for vår studie, ellers vil den noe høyere forekomsten gi jevnere fordeling, noe som igjen kan premieres med en høyere Kappa.

Om en sammenlikner våre Kappaberegninger med artiklene fra vår forskningsoppsummering, må vi konkludere med at vår interrater- reliabilitet fremstår som dårligere enn for de fleste sammenliknbare gjennomgangene, som i Lilford et als omfattende gjennomgang (37), der laveste der laveste Kappa målt var .32. Dette kan ha med publikasjonsbias å gjøre, det vil si at vårt

forskningsutvalg har høyere Kappa enn ikke- publisert forskning, eller at vi har et svært lite selektert utvalg som vil ha en lavere forekomst av unngåelige dødsfall enn i de fleste andre studier, og dermed stor skjevhet i dataene, som igjen gir lavere Kappa. Imidlertid er vår IRR for vurderinger av helsehjelpsproblemer, der forekomsten var 33% også relativt lav (k=.268) i sammenlikning med andre studier. Lilford et al. bemerker at reliabiliteten i deres studie ser ut til å være lavest der det er en implisitt eller ustrukturert metode, og det som undersøkes er årsakssammenhenger, mens reliabiliteten høyere der fokuset er på utkomme(37). Fokuset på årsakssammenhenger hos Lilford ser ut til å tilsvare prosessrelatert fokus hos Amalberti et al.(2011), og kan gjenkjennes i beskrivelsen av vår helsehjelpsproblemvariabel, der instruksen omtaler « *(…)ethvert tidspunkt der helsehjelpen falt under aksepterte standarder(…)»*, mens det også er et utkommefokus i den sammedefinisjonen ved at en skade må ha vært tilstede som følge av svikten (21). Metoden for vår gjennomgang inkluderer både eksplistte elementer, ved skåring av kategorier, og implisitte, ved at legegranskerne skal beskrive helsehjelpsproblemene i fritekst. Her ser vi at konsekvensene av det at vi ikke finner relevante studier der Gwets AC1 er beregnet til vår forskningsgjennomgang, er at sammenlikningsgrunnlaget mellom studiene blir dårligere ettersom fordelingene stort sett er skjeve og i større eller mindre grad antas å påvirke Kappa.

En uformell validitet ved vår studie kan en tenke seg følger av at reliabilitetseksperimentet speiler en vanlig klinisk situasjon, der leger eller pasienter benytter seg av en uavhengig «Second opinion». Legers vurderinger nyter stor tillit både innen helsetjenestene og i befolkningen, og en kan antageligvis utgå fra at denne er fortjent ved at de fleste pasienter opplever å bli møtt med oppriktighet og kompetanse, samt få hensiktsmessig diagnose og virksom behandling. Der to legespesialister uavhengig av hverandre gir den samme diagnosen, blir dette en slags klinisk «gullstandard». Variasjonen i et tenkt forsøk der de samme klinikerne skårer et subjekt i forutbestemte kategorier, er kanskje akseptabel fordi klinikerens virke i stor grad innebærer å gjøre slike vurderinger under en grad av usikkerhet, for eksempel i form av diagnosekoder eller behandlingsstrategier, og også vil la tvilen komme pasienten til gode eller i det minste handle etter ikke- skade prinsippet. Denne validiteten ser ut til å bestå av en erfarer at konklusjonene reproduseres, foruten legeprofesjonens gode rykte.

Med tanke på reproduserebarhet, vet vi at journalgjennomgangsmetoden er kjent for å produsere lav til moderat reliabilitet(30), det vil si at reproduserbarheten ofte er dårlig. En konsekvent dårlig reproduserbarhet kan, ,med bakgrunn i Bauer et als (2000) beskrivelse av reliabilitet innen den positivistiske forskningstradisjonen, tolkes som at en i realiteten måler et artefakt i forskningsprosessen heller enn et sant fenomen(13). Til forsvar for journalgjennomgang kan en kanskje argumentere for en legegranskers vurdering av hvorvidt faglige standarder er fulgt som oftest vil sammenfalle med faglige og lovmessige standarder, og at hva som er objektivt sant om

pasientens utkomme som følge av et helsehjelpsproblem og enigheten for disse vurderingene, ikke nødvendigvis er like relevant som funnene i seg selv, forutsatt at en kan nyttiggjøre seg disse til pasientenes beste.

Deler av skåringsprosessen kan kanskje beskrives som at legegranskerne utfører en type *innholdsanalyse* av tekster fra journaldokumentene, og beskrive validiteten av journalgjennomgang med kvalitative validitetsbegreper for denne typen analyse Krippendorff (1980). En kan se for seg at meningsinnholdet for noe abstrakte begreper som «unngåelighet» videre innehar et stort tolkningsrom, og ifølge Krippendorff (1980) vil dette kunne påvirke *konstruktvaliditeten*. Intensjonen bak det å benytte ekspertgranskere er å få frem skjønnsmessige vurderinger av journalene, men disse vurderingene må tilordnes, og en kan se for seg at de kanskje også må *tilpasses* et utfallsrom gitt av måleverktøyets validitet. I metodekapittel 5.3.3 vurderes to av våre variabler som enten ikke- uttømmende eller ikke gjensidig utelukkende, og dette kan også påvirke variablenes *konstruktvaliditet*. Popping (2019) beskrev reliabilitet som en bestanddel av *semantisk validitet(8),* og dersom vi utgår fra Krippendorffs (1980) definisjon av denne typen validitet, må vi vurdere om kategoriene reflekterer et meningsinnhold, ved at tvetydigheten og tolkningsrommet er snevret inn(43).

Dersom vi ser på instruksen for vår variabel *Helsehjelpsproblem*, ser vi at dette defineres som *"Et hvilket som helst tidspunkt der helsehjelpen til pasienten var under en akseptabel standard og førte til skade"*, og sannsynligheten som skal vurderes for å skåre et helsehjelpsproblem er *«et visst belegg*» (Vedlegg 3).

En kan si at instruksen for helsehjelpsvariabelen legger føringer for en vid definisjon og et stort tolkningsrom, og når kategorien som skåres beskriver at *«et visst belegg*» er tilstrekkelig for å skåre et helsehjelpsproblem, bidrar dette til inntrykket av at dette er en type «lavterskel» helsehjelpsproblem, der det vil være relativt lett å oppnå enighet, også ved stor usikkerhet, og en kan argumentere for at variabelens semantiske validitet, etter Krippendorffs (1980) definisjon, er lav. Resultatet kan være at en oppnår høy reliabilitet for en variabel med lav validitet, noe som gjør at enigheten en observerer ikke er like nyttig, fordi en ikke nødvendigvis måler et sant fenomen. Kort oppsummert kan en kanskje si at det er slike elementer ved skåringen de tilfeldighetskorrigerte koeffisientene skal kompensere for.

Fra Haywards (2007) metodeartikkel vet vi at det advares sterkt mot å generalisere funn til en større befolkning der en ikke kjenner spesifisiteten eller der denne er lav, ettersom dette kan gi en dramatisk overvurdering av forekomsten i befolkningen(44) En kan forvente fra instruksjonen og kategoribeskrivelsen for vår variabel *Helsehjelpsproblem* at denne vil ha høy sensitivitet og favner

mange av de journalene der det har forekommet helsehjelpsproblemer, men samtidig lav spesifisitet, og kunne gi *type 1- feil,* der en får falskt positive funn om helsehjelpsproblemer. Dette vil ha konsekvenser for overførbarheten av funnene for helsehjelpsproblemer.

Ved dikotome variabler som i vårt forsøk, kan en utgå fra at èn av granskerne tar feil og den andre har rett ved uenighet, uten at en kan vite hvem uten noen referanseverdi. Der granskerne er enige, er det kanskje mer sannsynlig at denne skåringen er den sanne verdien enn ikke, men det kan også være den samme feilkilden som påvirker begge og gir enighet for gal skår. Igjen må en lene seg på en noe uformell validitet, at legegranskerne i seg selv vil være valide skårere i omgang med dokumentasjonen, forutsatt at de er enige.

En kan i likhet med Popping (2019), gjøre seg ulike tanker om hvor grensen for en reel, velinformert *enighet* og en *tilfeldig enighet* som følge av for stor grad av gjetning går(8), og et naturlig spørsmål som oppstår vil være om forskeren i det hele tatt kan identifisere og håndtere gjetning på noen god måte. Hos legegranskeren, som er godt vant til å tolke journaldokumentasjon, også er erfaren i å ta «dikotome» avgjørelser med viss en grad av usikkerhet, kan en se for seg at det kan være en intensjon også bak hvilket utslag usikkerheten får. Som eksempel kan se for seg at en legegransker som tolker helsepersonells dokumentasjonsplikt strengt, vil skåre et helsehjelpsproblem i tvilstilfeller, dersom alvorlige mangler i journalen forårsaker usikkerheten.

Dersom en slik tendens er tilstede kan den i ytterste konsekvens være det Sedgwick og Greenwood (2015) kaller en *Hawthorne- effekt*(34), der granskerne ser noe av deres rolle i forsøket som å være «oppdragere» og også påpeker uhensiktsmessig praksis i journalene som ikke har ført til helsehjelpsproblemer. I skåringsinstruksen for PRISM2 blir det også vektlagt andre formildende og skjerpende omstendigheter som vil påvirke utslaget ved usikkerhet, som om det var aspekter ved pasientens helse som forverret utfallet av hendelsen, eller om det ble iverksatt hensiktsmessige tiltak da problemet ble oppdaget.

Dersom en slik Hawthorn- effekt er tilstede, kan også *nestenuhell* inngår i vår statistikk, noe som Amalberti et al (2009) mener ikke er ønskelig (21). Slike nestenuhell skal etter ehåndboken for OUS meldes (18), noe som inngår i det Amalberti et al. kaller et *prosessrelatert* og ikke *pasient- eller utkommerelatert* fokus(21).Om vi utgår fra at praksisen er tilsvarende ved større norske sykehus, vil begge granskerne i sin kliniske praksis være vant med at nestenuhell rapporteres.

Definisjonen for problemer i helsehjelpsproblemer for PRISM2 kan motvirke en slik effekt, ved at den består av èn *prosessrelatert* betingelse: «*Et hvilket som helst tidspunkt der helsehjelpen til pasienten var under en akseptabel standard…»* og èn *utkommerelatert* betingelse « *…og førte til skade*». En kan si derfor si at helsehjelp under en akseptabel standard i vår gjennomgang er en

nødvendig, men ikke tilstrekkelig betingelse for helsehjelpsproblem, og forutsatt at granskerne har et bevisst forhold til definisjonen, vil denne forhindre inklusjonen av nestenuhell.

Det bemerkes at *skade* ikke er definert i skåringsinstruksen, slik at kriterier for alvorlighet og varighet av en slik skade blir en skjønnsmessig vurdering. Dersom enighet er terskelen for deteksjon av slik skade, tilsier våre tall at så mye som 19.5% (n=39, kvinner *n*=15) av utvalget ble påført skade som følge av helsehjelpen, og uten enighetskravet er det høyeste funnet for menn, der gransker B finner slik skade i 44.8% av menns journaler. Tallene må sees i sammenheng med at sannsynlighetskravet for variabelen er «et visst belegg», og om en utgår fra at skade vil være lettere å identifisere i journalene enn helsehjelpsstandarden, kan en tenke seg at usikkerheten ved «et visst belegg» oftere vil omhandle vurderingene av helsehjelpsstandarden enn om det foreligger en skade, eventuelt at usikkerheten gjelder når skaden oppstod.

Flere våre variabler har kategoribeskrivelser som tar høyde for usikkerhet, også ved formuleringer som «mer enn et snev av», eller at det finnes kategorier som «umulig å bedømme». For den dikotome variabelen som konkluderer endelig om forebyggbart dødsfall, er definisjonen «mer enn 50% sannsynlig»; det vil si at det er bygget inn en usikkerhet som kan være så høy som 49%. Dette må tas inn i vurderingen av hva en konklusjon om forebyggbart dødsfall innebærer i vår studie. En kan se for seg at såpass vide definisjoner reduserer behovet for å korrigere for tilfeldig enighet med tilfeldighetskorrigerte koeffisienter, ettersom et stort tolkningsrom vil fange de fleste journaler der det er indikasjoner på dette.

I tillegg til at en 50% sannsynlighet for unngåelig dødsfall kan sies å ta høyde for en stor grad av usikkerhet, begrenses ikke disse vurderingene til pasienter som uten helsehjelpsproblemet ville overlevd sykehusoppholdet med tilstrekkelig helsehjelp. Haywards (2007) anbefalning om å fokusere på potensiell overlevelse ved optimal helsehjelp heller enn unngåelighet grunnet uønskede hendelser(44), kan synes å være hensiktsmessig dersom en ønsker å skille mellom disse pasientgruppene, og en kan se for seg at det også kan minke risikoen for Drors (2020) *lojalitets og minside*- bekreftelsesfeil, ved å flytte fokuset fra hvordan helsetjenesten *har* prestert ved helsehjelpsproblemer, til hvordan helsevesenet *burde ha* presteret. Her ser det ut til at Hayward et. als anbefaldninger er i tråd med Amalbertis (2011) syn på pasientsikkerhet, der en går fra et p*rosessfokus* til et pasient- og *utkommefokus*.

Et argument mot Haywards (2008) anbefaling, er at denne kan gi inntrykk av å utgå fra at unngåelige dødsfall enten ikke kan forekomme hos pasienter med dårlige overlevelsesprognoser, eller at det at å undersøke forebyggbare dødsfall for slike pasienter er uinteressant. Dette er også et

syn forfatteren ser ut til å bekrefte i sin studie Hayward og Hofer (2001) ved at palliative pasienter utelates fra utvalget(35). Dette vil kunne utgjøre et rettferdighetsproblem, dersom det fører til at denne delen av forskningen på pasientsikkerhet ikke kommer slike pasienter til gode. På en annen side kan en forstå at det er fristende, kanskje også hensiktsmessig, å dempe « støyet» i utvalget etter Hogans (2016) beskrivelse av at en ved deteksjon av unngåelige dødsfall leter etter et *svakt signal*(19).

Fra kapittel 2.1 kan en oppsummere at definisjonene av helsehjelpsproblemer innen forskningen varierer svært mye fra forfatter til forfatter når det gjelder hvilke prosesser eller pasientutfall som inkluderes. Hvor vidt definisjonene favner og hvilke hendelser som inngår i definisjonene vil legge føringer for hvordan granskeren leter etter og hva som tolkes som uønskede hendelser, og dermed påvirke enighet og dermed også fordeling og interrater-reliabilitetsberegningen for de målene som er tilgjengelige idag. En kan gjøre seg tanker om at det for helsepersonell vil ligge et tolkningsrom i de noe abstrakte begrepene «uønsket» hendelse, «tilstrekkelig» helsehjelpskvalitet eller «unngåelig» dødsfall, og at personlige verdier og erfaringer, samt syn på helsetjenestenes funksjon og ressurser påvirker dette. Profesjonelle normer kan også antas å variere fra spesialitet til spesialitet og fra et kollegialt felleskap til et annet, noe som kan tenkes å påvirke tolkningen.

Vi vet at sykehus er hierarkiske organisasjoner, og at vår studie ledes ved en avdeling på det sykehuset der sykehusoppholdene som skal granskes har vært, og at klinikerne som har hatt behandlingsansvar og skrevet journalene jobber ved det samme sykehuset. Den ene granskeren vår er ansatt ved et annet sykehus, og ble rekruttert som ekstern gransker, mens den andre er ansatt ved OUS i en leder- og forskerstilling (vedlegg 2). Denne granskeren vurderes å ikke være i et kollegialt felleskap med klinikere som har hatt behandleransvar for journalutvalget vårt.

Hovedstudien har tatt forhåndsregler for å unngå det Dror (2020) beskriver som *lojalitets- og minside- antagelser* som kan utsette ekspertgranskeren for bekreftelsesfeil, men lojalitetshensyn som følge av at legegranskerne vurderer et arbeid utført av andre leger kan ikke utelukkes eller kontrolleres for. Drors *snøball-kaskadeeffekt*, som kan oppstå der granskerne samarbeider(33), er i vår studie også unngått ved at granskerne har utført journalgjennomgangen uavhengig og på forskjellige steder og til forskjellige tider. Granskerne har heller ikke gjennomført felles kursing for forsøket utover skåringsinstruksen, og selv om dette er påpekt å kunne være en svakhet i reliabilitetsprotokoller og gi større variasjon i skåringene(8), kan en i teorien se for seg at kursing kan gi rom for felles bekreftelsesfeil, spesielt ettersom det ikke finnes retningslinjer for hva denne skal inneholde(47).

Våre granskere er klar over hensikten med forsøket, som for Hawthorne- effekten, slik den beskrives av Sedgwick og Greenwood (34), kan oversettes til vårt forsøk med at granskerne vurderes for sin evne til å identifisere helsehjelpsproblemer. En se for seg at om de samme legespesialistene gjennomgår de samme journalene i en mindre formell setting som klinikere, vil de ikke identifisere like mange helsehjelpsproblemer dersom Hawthorne effekten spiller inn i vårt forsøk[58]. Tversky og Kahnemanns (1974) *representativhetsprinsipp* (32) kan tenkes å gjøre seg gjeldende i vår studie for eksempel ved at en eller begge legegranskerne tar utgangspunkt i at multisyke og eldre oftere utsettes for helsehjelpsproblemer og dermed overvurderer forekomsten i et utvalg som har mange eldre, multisyke og hjelpeavhengige, eller at de overidentifisere helsehjelpsproblemer i journaler der de konkluderer med en generelt dårlig kvalitet på helsehjelpen eller underidentifiserer dem i opphold med generelt god helsehjelpskvalitet. En av Brennan et als (1991, 2004) konklusjoner fra deres journalgjennomgangsstudie var at mer alvorlige hendelser var knyttet til mer uaktsomhet (36, 61), og en ser for seg at også slike funn kan være påvirket av at *representativhetsprinsippet* spiller inn. En kan se for seg at slike bekreftelsesfeil vil gi bedre reliabilitet i de tilfellene de virker på granskerne på samme måte, og dårligere reliabilitet om de virker på granskerne «hver sin vei» eller kun påvirker den ene granskeren. Slike bedømmelsesfeil vil være svært vanskelig, om ikke umulig å kontrollere for i vår studie.

Tversky og Kahnemanns (1974) beskrivelser av bekreftelsesfeller som følger av *feiloppfatninger av tilfeldighet* (32) kan tenkes å ha innvirkning på granskerne dersom rekkefølgen av journaler ikke tilsier at helsehjelpsproblemer blir funnet med «jevn frekvens». Dersom det identifiseres forebyggbart dødsfall i to journaler på rad, vil en se for seg at granskeren kan ha større motstand mot å gjenkjenne dette også i den tredje journalen, eller at forekomsten en gransker finner i de første 100 journalene kan påvirke hvor høy forekomst granskeren er «villig» til å identifisere i de neste 100, selv om det er ikke er noen grunn til at helsehjelpsproblemer fordeler seg jevnt i journalutvalget. En kan også se for seg at dersom granskeren finner flere helsehjelpsproblemer innen èn spesialitet, avdeling eller pasientgruppe, uten at dette utgjør en systematisk skjevhet for utvalgsstørrelsen, kan *Gamblerens bekreftelsesfelle* gjøre at dette ikke oppleves tilfeldig eller «rettferdig» og gjøre at en ubevisst «leter grundigere» og dermed finne flere helshjelpsproblemer i påfølgende liknende journaler. Imidlertid kan en se for seg at det var nettopp tiltroen til tilfeldighetsmomentet ved nettopp myntkast som gjorde at *Gamblerens bekreftelsesfelle* gjorde seg gjeldende i Tversky og Kahnemanns forsøk, slik at dette ikke er en betydningsfull effekt i vårt forsøk.

[58] Scenariet er tenkt korrigert for bruken av skjemaverktøy i forsøket vårt

Dror (2020) erfarer også at ekspertens bekreftelsesfeil kan oppstå som følge av ubevisst bruk av *referansemateriell,* der granskeren ser bort i fra enkelte funn ettersom det for øvrig ikke passer til granskerens erfaringer(33).En kan se for seg at dokumentasjonskvaliteten kan påvirke gransker, ut ifra om det som inngår og er tilstrekkelig beskrevet i dokumentasjonen appellerer til den enkelte granskeren som *referansemateriell*.

Ved en nærmere undersøkelse av rådataene vår dokumentasjonskvalitetsvariabel, observeres det at gransker B har vurdert manglene ved èn journal som såpass alvorlige at det ikke var mulig å bedømme helsehjelpsproblemer, samtidig som samtidig skåret at det var umulig å trekke slutning om inngripende og uhensiktsmessige prosedyrer, og skåret at helsehjelpen av dårlig kvalitet. Samtidig har gransker A vurdert av helsehjelpen var av utmerket kvalitet, at journalene hadde *noen mangler* og at det ikke hadde forekommet inngripende og uhensiktsmessige prosedyrer. Ut ifra Drors teorier kan en her se for seg at slik uenighet kan oppstå som følge av at det hver enkelt gransker ubevisst bruker som *referansemateriell for* å identifisere helsehjelpsproblemer enten er til stede eller ikke for hver gransker i journalen og det dermed oppstår uenighet.

For variabelen dokumentasjonskvalitet ser en at gransker A har skåret alle journaler som tilfredsstillende for vår dikotomiserte variabel (vedlegg 6). En konsekvens av dette er at Kappa ikke kan beregnes, grunnet antagelsen om at det ikke kan være reliabilitet der det ikke er variasjon, som er en argumentasjon som støttes av Krippendorff (1980). Vi ser at samtidig tilsvarer kappa prosent enighet (98.4%) og ligger svært nær perfekt enighet. Det at Kappa ikke kan beregnes kan sees som en uheldig konsekvens av dikotomisering av variabelen, men det illustrerer uansett noe som må sees som en svakhet ved Kappa. Gransker B har skåret 3 journaler som utilfredsstillende, og dette utgjør en liten variasjon i datasettet, og selv om gransker B også hadde skåret alle som gransker A og det ikke er variasjon mellom dem, så kan en tenke, rent intuitivt og uten noen matematisk begrunnelse, at en reliabilitetskoeffisent som går fra -1 til 1, bør beregnes som 1 og reflektere 100% enighet, perfekt reliabilitet og ingen variasjon.

Dersom vi undersøker fordelingene fra vårt reliabilitetseksperiment nærmere, ser en at gransker B identifiserer flere «problematiske» journaler for alle variabler enn gransker A, og på overflaten ser ut til å være en noe «strengere» gransker, unntatt for variabelen *helsehjelpsproblemer* for gruppen kvinner, der gransker A identifiserer *et visst belegg i* noen flere journaler enn gransker B. Tidsbruksvariabelen (Vedlegg 6) kunne tenkes å få frem andre forskjeller ved granskernes skåringsprosess. Vi ser av analysene for denne at granskere har relativt lik tidsbruk for utvalget som helhet og for menn, og det eneste signifikante funnet var for kvinners journaler, der gransker A i gjennomsnitt brukte 4 min 43 sek mer tid enn gransker B per journal (vedlegg 6). Variasjonen i tidsbruk er imidlertid noe større for alle gransker Bs gjennomganger (vedlegg 6) Begge

legegranskerne bruker mer tid der det identifiseres helsehjelpsproblemer, og denne samvariasjonen er størst for gransker A, unntatt der det konkluderes med *unngåelig dødsfall* i gruppen menn, der den er størst for legegransker B (vedlegg 6).

Sett i sammenheng med granskernes tendens til å identifisere problemer i journalene, ser en at gransker A bruker noe mer tid enn gransker B på variabelen *helsehjelpsproblemer* for kvinners journaler, som er den eneste variabelen og gruppen gransker A også finner flere problematiske journaler for enn gransker B. Gransker A bruker imidlertid også mer tid på kvinners journaler for unngåelige dødsfall, selv om gransker B her finner flere problematiske journaler. Ettersom gransker B bruker mer tid enn A per unngåelige dødsfall denne granskeren finner for gruppen menn, og gransker b har identifisert den største mengden problemer i disse journalene uavhengig av variabel. Her kan en se for seg at gjennomgangen av unngåelige dødsfall for menn av ukjente grunner har vært mer tidskrevende for gransker B, eller også at den har vært mindre tidskrevende for gransker A enn de andre gjennomgangene.

En kan se for seg at tidsbruk også kan påvirkes av andre aspekter ved skåringsprosessen enn hvor vanskelige subjektene er å skåre eller hvor sannsynlig det er at det konkluderes med helsehjelpsproblemer, som at multisyke og kronisk syke ofte vil ha en mer omfattende og kompleks journal å gjennomgå, og at tidsbruk henger sammen med størrelsen på arbeidet ved gjennomgang av journalen hos hver pasient, heller enn at det er en direkte årsakssammenheng mellom unngåelig dødsfall og tidsbruk. Det vites heller ikke om helsehjelpsproblemer i seg selv fører til større mengder dokumentasjon, eller om det er høyere forekomst av helsehjelpsproblemer hos pasienter med omfattende journaler eller dårlig dokumentasjonskvalitet. En kan også se for seg at enkelte variabeltyper eller variabler med dårlig semantisk validitet, som ordinale variabler der kategoriene ikke har et tydelig meningsinnhold kan ta lengre tid å skåre. Hvor vide utfallsrommet for kategoriene er kan en også mistenke at vil påvirke dette, for eksempel ved at det går fortere å bestemme seg for kategorier som rommer alle journaler med « et snev av» eller « et visst belegg» enn kategorier som beskrives som «definitivt».

Kategorien et *visst belegg* for variabelen *helsehjelpsproblemer* skåres her mer liberalt av begge granskere enn tilsvarende «positivt funn»-kategorier for de andre variablene, og dette gir jevnere granskermarginaler. Selv om flere av variablene nærmer seg det Popping (2019) beskriver som *marginal homogenitet*(8), ved at granskernes fordeling over de samme kategoriene er nesten like, er *Helsehjelpsproblem* den eneste variabelen som ser ut til å nærme seg *uniform marginalfordeling*, der fordelingen over alle kategoriene er lik. Vi vet at Cohens Kappa «trives best» med jevne fordelinger, og i samsvar med dette, er det for denne variabelen Kappa beregnes høyest. Samtidig er det kun for denne variabelen at Gwets AC1 samtidig beregnes relativt langt under prosent enighet,

og gir kvalitative tolkninger som samsvarer med eller ligger i tilstøtende tolkningsintervall som Kappa.

Det mest kjente kappaparadokset, der en ser høy enighet og lav K ved skjeve fordelinger, ser en at forekommer hos de fleste av våre variabler, med forbehold om at Gwets AC1 til sammenlikning er et mer hensiktsmessig mål for slike fordelinger som påpekt av Gwet (2008) og Zhao, Liu og Deng (2013) samt Wongpakaran (2013) (41, 46, 52). Det følger imidlertid andre paradokser med målene, som først blir synlige der fordelingene forandrer seg fra ett måletidspunkt til et annet, og som vil være vanskeligere å påvise det er ett tidspunkt for alle skåringene. Flere forfattere fra vår gjennomgang presenterer svært tydelige eksempler der fordelinger er valgt for å provosere frem paradokser hos koeffisentene, og for ikke- statistikere kan en anta at mange av disse vil være vanskelige å identifisere der fordelingene og dermed utfallet av paradoksene ikke er like ekstreme (41, 46, 50).

Helsehjelpsproblemer beskrives i Hogans (2016) fagartikkel som sjeldne hendelser(19), og dette stemmer med samtlige funn fra utvalgt forskning i kapittel 2.5, og også vår studie når det gjelder unngåelig dødsfall. Vi forventet skjeve fordelinger, og at Cohens Kappa også kunne rammes av paradokser, noe som ser ut til å ha inntruffet. Gwets AC1 bemerkes å kunne være blant de mer liberale koeffisientene ved skjeve fordelinger og ligge nær opp mot prosent enighet(46), og dette observeres også i vår studie. En kan ikke utgå fra at Gwets AC er en referanseverdi for Kappa, og en må holde i mente at AC1 også kan være en urimelig liberal koeffisent, som får Kappa i sammenlikning til å fremstå som mer konservativ enn den er.

Selv om Kappas generelle svakheter påpekes av svært mange forfattere, gjør de samme forfatterne gjerne konklusjoner på bakgrunn av beregnet Kappa uten å undersøke hvordan Kappa påvirkes av datafordelingen. Det er overraskende at Gwets AC eller andre mål[59] som er utviklet for å håndtere kappaparadoksene, ikke benyttes i større grad. Det ene paradokset som oppstår for prosent enighet er velkjent og kan forstås av ikke- statistiskere, og kan for våre beregninger se ut til å være et omtrent like anvendelig mål Gwets AC1 ved svært skjeve fordelinger, ettersom disse vil gi samme kvalitative tolkning. En kan gjøre seg tanker om prosent enighet som kanskje et primitivt, men kanskje også mer «ærlig» mål på, ettersom paradokset der tilfeldig skåring kan være reliabelt er velkjent og lettere å forstå enn de tilsynelatende noe uberegnelige og uintuitive paradoksene som følger IRR- koeffisientene. Med bakgrunn av forståelsen fra måleteori i kapittel 3.1, kan en kanskje også argumentere for at en større feilkilde som kan forstås og til en viss grad beskrives, er bedre enn feilkilde av ukjent størrelse som ikke forstås fullt ut av forskeren. En overføring av Prosent enighet

[59] Som PABAK, prevalence adjusted, bias adjusted kappa.

forutsetter imidlertid at en har en rigid protokoll og at subjektene er lette å skåre, ifølge Zhao Liu og Deng (2013).

Gwets 3 hovedpunkter tegner opp svært stramme rammer for generaliserbarheten for vårt forsøk, og om en trekker inn Hayward et als(44) problematisering av generalisering av resultater fra metoder som kjent å produsere lav til moderat reliabilitet der en ikke kjenner spesifisitet eller sensitivitet, må en konkludere med at en bør en være svært forsiktig med å overføre reliabiliteten fra vårt forsøk til andre forskningssituasjoner eller generalisere forekomsten av helsehjelpsproblemer som fremkommer i vårt forsøk til en større befolkning.

Gwet (2019) skal en for overføringsverdien av IRR fra reliabilitetsstudier vurdere om en kan anta at andre granskere vil enes for det samme utvalget og at de samme granskerne vil opprettholde enigheten om de skårer andre subjekter og på bakrunn av dette konkludere med om enigheten «hører til» dette ene eksperimentet når akkurat disse granskerne skårer disse subjektene. Gwet påpeker at konklunsjonen av denne vurderingen svært ofte vil vær at overføringsverdien er lav , ettersom forutsetningene om at granskerne og subjektene er gode representanter for sine populasjonen og at protokollen har høy grad av rigiditet sjelden sammenfaller(41)) Det at begge deler skal begrunnes kan påvirke

En kan gjøre seg forskjellige tanker om hva en i realiteten måler ved interrater- reliabilitet. En lav interrater- reliabilitet kan innebære et stort antall subjekter som er vanskelige å skåre, mangler ved dokumentasjonsgrunnlaget, validitetsproblemer ved skjemaverktøyet eller at granskerne ikke er reelt utskiftbare, påvirkes av bekreftelsesfeil eller at det er forskjeller i hvordan de forholder seg til skåringsinstruksjonene eller kursingen. Interrater- reliabilitet fremstår etter en slik oppsummering som et svært komplekst og kanskje også lite konsist mål, selv der granskerne skårer alle subjektene samvittighetsfullt.
Både kappa og Gwets AC skal være egnet til å validere skåringsinstrumenter, men er sårbare for antall kategorier og andre konstruktsaspekter som kan påvirke fordeling(46). Det vil si de er avhengige av instrumentets reliabilitet for å vurdere dets reliabilitet, og følger *sirkulær logikk* ifølge Zhao, Liu og Deng (2013) (46). I vårt forsøk er dette svært alvorlig svakhet, ettersom vi ønsker å bidra til å validere et skjema der intern konsistens og intra-rater reliabilitet ikke er undersøkt i forkant. Dette kan ha påvirket fordelingen uten at vi vet om den samme fordelingen vil reproduseres om verktøyet brukes i andre forsøk, noe som kan gjøre at IRR- målene i seg selv ikke vil kunne reproduseres.

Kanskje kan en også tenke seg at tilfeldig skåring kan øke med et relativt stort utvalg på 200 journaler, ettersom skåringsoppgaven etter hvert kan oppleves monoton[60].

Å tvinge en viss formalisering av de ukjente kognitive prosessene ved at granskere visualiserer og skårer en tallskala som hos Hayward og Hofer(35), kan antageligvis være nyttig kompensere for semantiske validitetsproblemer der kategoriene i et skåringsinstrument har tvetydighet eller et stort tolkningsrom

8.0 Konklusjoner

1: Hva er forekomsten av uønskede hendelser og forebyggbare dødsfall for utvalget, og i hvor mange journaler er det enighet for disse vurderingene? ¨

Ved enighet mellom granskerne ble det funnet et *visst belegg* for helsehjelpsproblemer i 33% (n= 66) av journalene. Kvinners journaler utgjorde 22.7% av disse (n=15).

Det ble konkludert med over 50% sannsynlighet for unngåelig dødsfall i 2% av journalene (n=4), og Kvinners journaler utgjorde 25% av disse (n=1).

2: Hva er interrater- reliabilitet målt ved Prosent enighet, Cohens Kappa og Gwets AC1 for variabelutvalget?

Prosent enighet for *helsehjelpsproblemer* var 86.2%, Kappa var .268 (KI .128- .406) og AC1 var.384 (KI .209- .488). Begge koeffisientene er innen Landis og Kochs (1974)(53) kvalitative tolkning *Fair* (IRR .21- .40).

Prosent enighet for *unngåelig dødsfall* var 88.5%, Kappa var .209 (KI -.006-.424) og innen tolkningen *Fair* (IRR .21-.40). Gwets AC1 var .865 (KI .806 - .925) og innen tolkningen *Almost Perfect (*IRR .81 – 1).

Enighet for *helsejelpsproblemer* var høyere for menns journaler (Prosent enighet 70.6%AC1 .400 (Fair) (KI 215-.586) / Kappa .358, (Fair) (KI .244-.471)) enn kvinners journaler (Prosent enighet 61.7%, AC1 .291 = Fair, (KI .078-.506)/ Kappa .172= *Slight*, (KI -.035-.379))

[60] . Dette vil omfattes av begrepet *granskerfatigue, ved lesing av artikkelsammendrag finner en* små effekter av, og ingen sammenliknbare studier for denne oppgaven (Se for eksempel Ling G, Mollaun P (2014).

Enighet var for *unngåelig dødsfall* var høyere for menns journaler (Prosent enighet 89.2%, AC1.872 = *Almost Perfect*, (KI .793-.952) /Kappa .305= Fair, (KI .244-.471) enn for kvinner (Prosent enighet 77%, AC1 .857= *Almost Perfect*, (KI .767-.947) / Kappa .088 = *Slight*, (KI -192-.368))

1. *Hvordan kan forskjellene i enighet og variasjon mellom våre interrater- reliabilitetsmål forkalres?*

Generelle tendenser for våre IRR- mål samsvarer med forventningene om hvordan målene påvirkes av datafordelingen etter gjennomgangen i Kapittel 4.
Gwets AC1 var høyere enn Cohens Kappa for alle variabler, og kun innen samme kvalitative tolkningskategori for *snev av belegg for helsehjelpsproblemer*, som har den jevneste datafordelingen i vårt variabelutvalg. For de øvrige variablene kan forskjellene beskrives som dramatiske for den kvalitative vurderingen og konklusjonen om reliabilitet. For den dikotomiserte variabelen *dokumentasjonsgrunnlag* kunne ikke Kappa beregnes, ettersom alle journalene ble skåret i èn kategori av den ene granskeren. For den samme variabelen ga AC1 og Prosent enighet svært høy reliabilitet og enighet. Cohens Kappa var ikke signifikant for *Inngripende og uhensiktsmessige prosedyrer.*

Unntatt for den relativt jevnt fordelte variabelen *problemer i helsehjelpen*, ga AC1 reliabilitet innen tolkningen *Almost Perfect* for alle variabler for utvalget som helhet, og lå nært opp mot prosent enighet. AC1 utgår fra fullstendig oppriktighet og variabel tilfeldig enighet, som kan synes å være rimelige antagelser med en rigid protokoll og motiverte granskere.

8.1 Begrensninger og anbefalninger for videre forskning

Det påpekes av blant andre Zhao, Liu og Deng (2013) at forskning med beregning av interrater-reliabilitet kan være et utvalg med urimelig høy IRR som følge av publikasjonsbias(46). Ettersom Kappa er mest brukt, hevder Zhao, Liu og Deng at det dessuten kan være overrepresentasjon av studier med mer jevne fordelinger og høyere forekomst av helsehjelpsproblemer, ettersom Kappa tendenserer til å være urimelig lav ved lav forekomst(46). Vi må for vår gjennomgang konkludere med relativt lav Kappa sammenliknet med det som gjenfinnes i vårt utvalg av forskningsartikler, og det er mulig denne konklusjonen er preget av at det foreligger publikasjonsbias.

Gwets AC1 kan ifølge Zhao, Liu og Deng derimot være for liberal(46) ved skjeve fordelinger, og rent intuitivt kan en mene at det er urimelig at AC1 skal ligge såpass tett opp mot Prosent enighet som den gjør i vår studie, dersom hensikten med koeffisienten er å korrigere for tilfeldig enighet og en annerkjenner at tilfeldig enighet forekommer. Kappa kan her også fremstå urimelig lav for vår studie i sammenlikning med det som i relaliteten kan være en urimelig høy Gwets AC1.

Det er ikke undersøkt om reliabiliteten ble dårligere utover i journalgjennomgangen, men 200 journaler er et omfattende materiale og en kan ikke gå ut ifra at presisjonen i vurderingene var like god for alle journaler.

Det at ingen pasienter er ekskludert fra vårt utvalg kan gi lavere reliabilitet, dersom det antas at for eksempel faglige standarder for psykiatri eller psykiatriske journaler er vanskeligere å skåre for legegranskere. Vi vet heller ikke om journalgjennomgang er like godt egnet til deteksjon av helsehjelpsproblemer for alle inkluderte pasientgrupper.

Våre data avdekker ikke om de som rammes av unngåelig dødsfall er pasienter som med tilfredsstillende helsehjelp ville overlevd sykehusoppholdet, og utvalget utelukker deteksjon av helsehjelpsproblemer som ikke ga dødelig utfall, noe som gjelder flertallet av befolkningen. For forskning der det er relevant å skille mellom disse pasientgruppene, kan det anbefales å inkludere en variabel som vurderer overlevelse med tilfredsstillende helsehjelp i en periode etter utskrivelse.

En ser en del lavere enighet for helsehjelpsproblemer og unngåelige dødsfall for kvinne, og multivariabel analyse for kjønn for å avdekke om denne sammenhengen er signifikant kan være interessant for fremtidige studier.

Som beskrevet i metodedelen (kap. 5.3.3) har det å kollapse variabel 16 gitt en dobling av forekomsten av unngåelig dødsfall, bedring i Prosent enighet og antatt bedre IRR[61]. Vi vet at IRR påvirkes av målenivå, og det er også påpekt semantiske validitetsproblemer ved kategoriene (Kapittel 5.3.3), som vi kan ha «dekket over» ved dikotomisering. Dette metodevalget anses å ha ført til informasjonstap fra den opprinnelige variabelen og gitt kunstig god reliabilitetsvurdering av variabelen, og dette er en vesentlig begrensning for hovedstudiens delmål 1, der målsetningen er å validere PRISM2- skjemaet.

Sirkulær logikk-paradokset fra kapittel 4.2.5 illustrerer en vesentlig begrensning ved våre mål for IRR, ved at målene selv påvirkes av reliabiliteten til konstruktet de skal beregne reliabilitet for, og det at intrarater og intern konsistens heller ikke er undersøkt, kan bidra til å gjøre dette til en begrensning for reliabilitetsberegningene. Øvrige problematiske aspekter ved våre koeffisienter kan oppsummeres som at Kappa et svært konservativt IRR- mål ved skjeve fordelinger, mens AC1 antageligvis kan gi for liberale beregninger og selv om effekten av dette er ukjent, kan det se ut som om dette har stor betydning for våre konklusjoner.

[61] IRR er ikke beregnet for den ordinale variabelen

Det er ikke satt en cutoff for IRR for vår studie, og det er ikke bestemt hvilken kvalitativ tolkning etter Landis og Koch (1977) som utgjør god nok IRR, eller hvilken koeffisient som skal følges. Dette må gjøres uavhengig av hvilket IRR- mål som velges, før resultatene kan vurderes for overføring til den større befolkningen som døde ved OUS i 2014.

Journalgjennomgang er en metode med store forventninger til legegranskernes evner til å beregne sannsynlighet under stor usikkerhet, og med de store utviklingene innen maskinlæring og kunstig intelligens, er det ikke gitt at mennesker i fremtiden er best egnet til denne oppgaven. Med et innslag av fremtidsoptimisme kan en avslutningsvis sende en varm tanke til en fremtid der teknologien kan avlaste granskere, ved for eksempel å gjøre strategiske journalutvalg basert på triggere, slik at journalgjennomgangsarbeidet blir mindre omfattende.

9.0 Litteraturliste

1. Rall M. To err is human - a summary of the IOM-Report. European journal of anaesthesiology. 2000;17(8):520.

2. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care. 2008;17(3):216-23.

3. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gurp PJ, de Vet HCW, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. BMJ Open. 2016;6(8):e011078.

4. NKF. Rapport for Nasjonal Journalundersøkelse med Global Trigger Tool 2011 10.10.2021. Available from: https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2013/rapport-for-nasjonal-journalundersokelse-med-global-trigger-tool-2011.pdf.

5. Rogne T, Nordseth T, Marhaug G, Berg EM, Tromsdal A, Sæther O, et al. Rate of avoidable deaths in a Norwegian hospital trust as judged by retrospective chart review. BMJ Qual Saf. 2019;28(1):49-55.

6. Flaatten H, Brattebø G, Alme B, Berge K, Rosland JH, Viste A, et al. Adverse events and in-hospital mortality: an analysis of all deaths in a Norwegian health trust during 2011. BMC Health Services Research [Internet]. 2017 10.3.2022; 17(1):[465 p.]. Available from: https://doi.org/10.1186/s12913-017-2417-7.

7. Helsedirektoratet. Lovdata. Lovdatano [Internet]. 2017 mars 13. Available from: https://lovdata.no/static/ROO/is-2017-2620.pdf.

8. Popping R. Introduction to Interrater Agreement for Nominal Data. Cham: Springer International Publishing; 2019. Available from: https://linkspringer.com.ezproxy.uio.no/book/10.1007/978-3-030-11671-2.

9. Lindahl AK, Håheim LL. Helsetjenesteforskning og helsetjenestenes kvalitet. tidsskriftetno [Internet]. 2017 3 21. Available from: https://tidsskriftet.no/2017/03/kommentar-og-debatt/helsetjenesteforskning-og-helsetjenestens-kvalitet.

10. Alfsen G, Lyckander L, Eng H, Lindboe A. Quality control of deaths by pathologists: A systematic approach for improving death certificate completions, notification procedures and death statistics. Archiv für Pathologische Anatomie und Physiologie und für Klinische Medicin. 2013;463:105-.

11. SINTEF. Håndbok for journalgjennomganger. Norsk pasientregister [Internet]. 2006 15.10. 2021; A679. Available from: https://www.sintef.no/globalassets/upload/helse/rapporter-npr_pafi/a679-handbok-journalgjennomgang.pdf

12. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ : British Medical Journal. 2015;351:h3239.

13. Bauer M. Classical content analysis: A review. Qualitative Researching with Text, Image and Sound: A Practical Handbook. 2000:131-51.

14. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf. 2012;21(9):737-45.

15. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ. 2015;351:h3239-h.

16. Hogan H. Preventable Incidents, Survival and Mortality

Study 2 (PRISM) - Medical Record Review Manual 2014 [Available from: https://www.england.nhs.uk/wp-content/uploads/2021/07/PRISM-2-Manual.pdf.

17. Spath P. Error reduction in health care : a systems approach to improving patient safety. San Francisco; Chicago, IL: Jossey-Bass ; AHA Press; 2000.

18. Reynard JRJSP. Practical patient safety. Oxford; New York: Oxford University Press; 2009.

19. Hogan H. The problem with preventable deaths. BMJ Quality & Safety. 2016;25(5):320-3.

20. Schwendimann R, Blatter C, Dhaini S, Simon M, Ausserhofer D. The occurrence, types, consequences and preventability of in-hospital adverse events - A scoping review. BMC Health Serv Res. 2018;18(1):521-9.

21. Amalberti R, Benhamou D, Auroy Y, Degos L. Adverse events in medicine: Easy to count, complicated to understand, and complex to prevent. Journal of Biomedical Informatics. 2011;44(3):390-4.

22. Universitetssykehus O. Uønskede hendelser, risikoforhold og forbedringsforslag i Achilles Oslo: Elektronisk håndbok; 2022 [Available from: Uønskede hendelser, risikoforhold og forbedringsforslag i Achilles.

23. Helsedirektoratet. Helsepersonelloven : med forskrifter og Helsedirektoratets kommentarer2019. Available from: https://lovdata.no/dokument/NL/lov/1999-07-02-64.

24. Helsedirektoratet. Lov om Spesialisthelsetjenesten 2021 [Available from: https://lovdata.no/dokument/NL/lov/1999-07-02-61.

25. Sari AB-A, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. 2007;334(7584):79.

26. Hibbert PD, Molloy CJ, Hooper TD, Wiles LK, Runciman WB, Lachman P, et al. The application of the Global Trigger Tool: a systematic review. Int J Qual Health Care. 2016;28(6):640-9.

27. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.

28. Lombard M, Snyder-Duch J, Bracken CC. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. Human communication research. 2002;28(4):587-604.

29. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Statist Med. 2002;21(14):2109-29.

30. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.

31. Bennet EM, Alpert R, Goldstein AC. Communications Through Limited-Response Questioning. Public Opinion Quarterly. 1954;18(3):303-8.

32. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974;185(4157):1124-31.

33. Dror IE. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. Analytical Chemistry. 2020;92(12):7998-8004.

34. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ : British Medical Journal. 2015;351:h4672.

35. Hayward RA, Hofer TP. Estimating Hospital Deaths Due to Medical Errors: Preventability Is in the Eye of the Reviewer. JAMA. 2001;286(4):415-20.

36. Brennan TA. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. Quality & safety in health care. 2004;13(2):145-51.

37. Lilford R, Edwards A, Girling A, Hofer T, Di Tanna GL, Petty J, et al. Inter-rater reliability of case-note audit: a systematic review. Journal of Health Services Research & Policy. 2007;12(3):173-80.

38. Deilkås ET, Bukholm G, Lindstrøm JC, Haugen M. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. BMJ Open. 2015;5(12):e008576-e.

39. Deilkås ECT, Risberg MB, Haugen M, Lindstrøm JC, Nylén U, Rutberg H, et al. Exploring similarities and differences in hospital adverse event rates between Norway and Sweden using Global Trigger Tool. 2017.

40. Schiøler T, Lipczak H, Pedersen BL, Mogensen TS, Bech KB, Stockmarr A, et al. [Incidence of adverse events in hospitals. A retrospective study of medical records]. Ugeskr Laeger. 2001;163(39):5370-8.

41. Gwet KL. Handbook of inter-rater reliability : the definitive guide to measuring the extent of agreement among raters. 5th ed. Gaithersburg, MD2021.

42. Laake P. Epidemiologiske og kliniske forskningsmetoder. Oslo: Gyldendal akademisk; 2007.

43. Krippendorff K. Content analysis an introduction to its methodology. Beverly Hills: SAGE; 1980.

44. Hayward RA, Heisler M, Adams J, Dudley RA, Hofer TP. Overestimating outcome rates: statistical estimation when reliability is suboptimal. Health Serv Res. 2007;42(4):1718-38.

45. Tinsley HEA, Weiss DJ. 4 - Interrater Reliability and Agreement. In: Tinsley HEA, Brown SD, editors. Handbook of Applied Multivariate Statistics and Mathematical Modeling. San Diego: Academic Press; 2000. p. 95-124.

46. Zhao X, Liu J, Deng K. Assumptions behind Intercoder Reliability Indices. Annals of the International Communication Association. 2013;36.

47. Kottner J, Gajewski BJ, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS). International Journal of Nursing. 2011;48(6):659-60.

48. Carter RELJDECRE. Rehabilitation research : principles and applications. 2016.

49. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46.

50. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9.

51. ten Hove D, Jorgensen TD, van der Ark LA. Interrater Reliability for Multilevel Data: A Generalizability Theory Approach. Psychol Methods. 2021.

52. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Medical Research Methodology. 2013;13(1):61.

53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

54. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

55. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010;78(1):40-64.

56. Rubin LH, Witkiewitz K, Andre JS, Reilly S. Methods for Handling Missing Data in the Behavioral Neurosciences: Don't Throw the Baby Rat out with the Bath Water. J Undergrad Neurosci Educ. 2007;5(2):A71-A7.

57. Kunnskapsdepartementet. Lov om organisering av forskningsetisk arbeid (forskningsetikkloven) 2017 [Available from: https://lovdata.no/dokument/NL/lov/2017-04-28-23.

58. Worldometers.com. Norway population (2020 and historical) 2022 [Available from: https://www.worldometers.info/world-population/norway-population/.

59. Kalseth J, Halvorsen T. Relationship of place of death with care capacity and accessibility: a multilevel population study of system effects on place of death in Norway. BMC Health Services Research. 2020;20(1):454.

60. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ. 2015;351:h3239-h.

61. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of Adverse Events and Negligence in Hospitalized Patients: Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.

Vedlegg 1 Personvernombudets tilråding

Oslo universitetssykehus

Oslo universitetssykehus HF

Postadresse:
Postboks 4950 Nydalen
0424 Oslo

Sentralbord:
02770

Org.nr:
NO 993 467 049 MVA

www.oslo-universitetssykehus.no

PERSONVERNOMBUDETS TILRÅDING

Til:	Mette Utheim
Kopi:	Trine Sand Kaastad
Fra:	Personvernombudet ved Oslo universitetssykehus
Dato:	17.09.2021
Offentlighet:	Ikke unntatt offentlighet
Saksnummer:	20/12111

«Pasientdødsfall i Oslo universitetssykehus vurdert ved retrospektiv journalgjennomgang (RCRR)»

Interrater- reliabilitet for hovedkonklusjoner fra legegranskeres journalgjennomgang ved dødsfall i sjukehus: En reliabilitetsstudie av skjemaverktøyet PRISM 2 for retrospektiv journalgjennomgang

Formål:
Formålet for masteroppgaven er å bidra til validering av skjemaverktøyet PRISM2 samt beskrive forekomsten av forebyggbare dødsfall og uønskede hendelser i utvalget.
PRISM 2 er oversatt til norsk etter godkjente kriterier (Kaastad, T.S 2015).
Formålet oppnås ved å beregne interrater- reliabiliteten for hovedkonklusjonene samt et utvalg relevante variabler fra verktøyet, samt å tallfeste begge legegranskeres konklusjoner og beskrive forekomsten av disse i de to datasettene.
Studien tar utgangspunkt i datamateriale fra to uavhengige journalgjennomganger med et mye brukt skjemaverktøy, og kan gi bedre forståelse for informasjonen overvåkningsverktøy som dette produserer. Dette vil inngå i forebedringskunnskapen i pasientsikkerhetsarbeidet ved OUS.

Oppgaven inngår i den pågående studien "Pasientdødsfall i Oslo Universitetssykehus vurdert ved retrospektiv journalgjennomgang (RCRR), (OUS sak 2015/6035).

Med hjemmel i forordning (EU) nr. 2016/679 (generell personvernforordning) artikkel 37, er det oppnevnt personvernombud ved Oslo Universitetssykehus (OUS).

Den dataansvarlige skal sikre at personvernombudet på riktig måte og i rett tid involveres i alle spørsmål som gjelder vern av personopplysninger, jf. artikkel 38. Artikkel 30 pålegger OUS å føre oversikt over hvilke behandlinger av personopplysninger virksomheten har. Behandling av personopplysninger meldes derfor til sykehusets personvernombud.

Før det foretas behandling av helseopplysninger, skal den dataansvarlige rådføre seg med personvernombudet, jf. personopplysningsloven § 10. Ved rådføringen skal det vurderes om behandlingen vil oppfylle kravene i personvernforordningen og øvrige bestemmelser fastsatt i eller med hjemmel i loven her. Rådføringsplikten gjelder likevel ikke dersom det er utført en vurdering av personvernkonsekvenser etter personvernforordningen artikkel 35.

Databehandlingen tilfredsstiller forutsetningene for melding etter forordning (EU) nr. 2016/679 (generell personvernforordning) artikkel 30.

Helseopplysninger kan behandles uten samtykke dersom behandlingen er nødvendig for formål knyttet til vitenskapelig forskning og samfunnets interesse i at behandlingen finner sted, klart overstiger ulempene for den enkelte, jf. personopplysningsloven § 9, jf. generell personvernforordning artikkel 6 nr. 1 bokstav e) og artikkel 9 nr. 2 bokstav j).

Personvernombudet har vurdert at den planlagte databehandlingen er nødvendig for kvalitetssikring av helsehjelpen, jf. pasientjournalloven § 6, annet ledd.
Bruk av helseopplysninger skal skje i samsvar med taushetspliktreglene, jf. helsepersonelloven § 26.

Mette Utheim har lovlig tilgang til rådata via sensitivt k- område som ansatt ved OUS og prosjektmedarbeider ved studien.

Personvernombudet tilrår at databehandlingen gjennomføres.

1. Oslo universitetssykehus HF ved adm. dir. er dataansvarlig virksomhet.
2. Avdelingsleder eller klinikkleder ved OUS har godkjent databehandlingen.
3. Databehandlingen skjer i samsvar med og innenfor det formål som er oppgitt i meldingen.
4. Data lagres som oppgitt i meldingen og i samsvar med sykehusets retningslinjer.
5. Oppslag i journal med formål å identifisere potensielle deltagere til studien gjøres av ansatte ved sykehuset som har selvstendig lovlig grunnlag for oppslaget. Se http://ehandboken.ous-hf.no/.
6. Eventuelle fremtidige endringer som berører formålet, utvalget inkluderte eller databehandlingen må forevises personvernombudet før de tas i bruk.
7. Publisering i tidsskrift forutsettes å skje uten at deltagerne kan gjenkjennes, hverken direkte eller indirekte.

Med hilsen

Tor Åsmund Martinsen
Personvernombud

Oslo universitetssykehus HF
Direktørens stab | Personvern

Personvernombudets tilråding

Oslo universitetssykehus

Oslo universitetssykehus HF

Postadresse:
Postboks 4956 Nydalen
0424 Oslo

www.oslo-universitetssykehus.no

trine.sand.kaastad@ous-hf.no

Vurdering av dødsfall
PROTOKOLL

Stab Pasientsikkerhet og Kvalitet
Trine Sand Kaastad
Mobiltelefon: +47 92 42 37 01

PASIENTDØDSFALL I OSLO UNIVERSITETSSYKEHUS VURDERT VED RETROSPEKTIV JOURNALGJENNOMGANG (RCRR)

Forekomst og forebyggbarhet av dødsfall knyttet til uønskede hendelser – validering og sammenligning av etablert verktøy

Pasientsikkerheten er helt sentral for all sykehusdrift, og det er innført landsdekkende systemer og metoder med tanke på å overvåke denne sikkerheten, bl.a. i regi av Kunnskapssenteret i form av Global Trigger Tool (GTT) og 30 dagers overlevelse (1, 2). I Oslo universitetssykehus (OUS) oppleves det et behov for å supplere disse nasjonale tiltakene for å få et mer nyansert bilde av kvaliteten i egen virksomhet med tanke på mulig læring og forbedringer. Relevant forskning om uønskede hendelser og dødsfall i sykehus er derfor vurdert med tanke på å finne godt egnede metoder for gjennomgang av egne pasientgrupper som kan gi valide slutninger.

Harvard Medical Practice Study

En gruppe på Harvard sto på 1980-tallet for omfattende studier av forekomsten av uønskede hendelser i en rekke New York State-sykehus (3, 4). Metoden man brukte - retrospektiv, strukturert, implisitt gjennomgang av pasientjournaler (RCRR) - ble bl.a. evaluert med tanke på enighet blant legegranskere i vurderinger av uønskede hendelser (5). I 12,9 % av tilfellene (971 av 7533) var to uavhengige legegranskere helt uenige om at det hadde forekommet en uønsket hendelse eller ikke, hvilket var flere enn de tilfeller der begge var enige om at det hadde vært en uønsket hendelse (757, 10 %). Dette er i tråd med funn fra en studie om vurdering av kvaliteten på helsehjelp i en indremedisinsk avdeling ved fagfellevurdering, der de eneste faktorer som nådde et enighetsnivå som ga grunnlag for sammenligning av avdelinger, var 'samlet kvalitet av helsehjelp' og 'vurdering av forebyggbarhet av dødsfall' (6). Goldman rapporterte i samme periode om en gjennomgang av studier gjort vedrørende fagfellevurderinger av kvalitet på helsehjelp (7), og etterlyste en forbedring av metoden med tanke på pålitelighet. Lilford et al oppdaterte og utvidet gjennomgangen 15 år senere (8), og de fant bedre enighet der det var eksplisitte kriterier for gjennomgangene enn ved implisitte (strukturerte) vurderinger. De foreslo derfor at en måte å få det beste ut av journalgjennomganger kunne være å bruke et sett med eksplisitte spørsmål kombinert med et avsluttende implisitt bidrag (8).

Nederland

Metoden fra Harvard Medical Practice Study (4) har blitt videreutviklet, bl.a. av grupper i Canada (9) og Nederland (10). I den nederlandske studien av uønskede hendelser og potensielt forebyggbare dødsfall fra 2009, gjennomgikk man et utvalg av pasienter der 50% var døde (10). Zegers et al undersøkte også om det førte til bedre inter-rater enighet om vurderingene ble gjort av to fagfeller fremfor en, hvilket de ikke kunne bekrefte

(11). Dette stemmer med funn ti år tidligere av Hofer et al (12), som konkluderte med at diskusjoner mellom granskere ikke økte påliteligheten av fagfellevurderinger av sykehuskvalitet.

England

Hogan et al. sin retrospektive studie av forebyggbare dødsfall som skyldtes problemer i helsehjelpen i engelske akuttsykehus (13), var basert på metodene i den nederlandske studien (14), i tillegg til i metodikken fra en dødsfallsstudie av Hayward & Hofer (15). Hogan et al. studerte utelukkende pasienter utskrevet fra sykehus som døde, i alt 1000, og introduserte begrepet 'problems in care' (problemer i helsehjelpen) der man tidligere har brukt uønskede hendelser. De konkluderte at problemer i helsehjelp i sykehus representerer en stor byrde, men at forekomsten av dødsfall som hadde 50 % eller større sjanse for å være forebyggbare 'bare' var 5,2%. For å vurdere inter-rater pålitelighet, ble 25 % av alle dødsfall gjennomgått av to leger, ellers nøyde de seg med en generelt kyndig spesialist (13). De fant moderat overensstemmelse i vurderingene av om det var 'problemer i helsehjelpen' og 'forebyggbart dødsfall' med κ-verdier på hhv 0.54 og 0.49.

Skandinavia

Tilsvarende gjennomganger for å avdekke forekomst av problemer i helsehjelpen generelt eller spesifikt knyttet til mortalitet i sykehus, er ikke publisert i Norge, mens man både i Danmark (16) og Sverige (17) har gjort studier av forekomst av uønskede hendelser i sykehus med utgangspunkt i RCRR-metoder lignende de nevnt over.

I Danmark har man i tillegg tatt i bruk en tilnærming, opprinnelig fra Institute for Healthcare Improvement (IHI) i USA, der helsepersonell vurderer 50 konsekutive dødsfall (18, 19). En rapport fra Dansk Selskab for Patientsikkerhed i 2013 (20) omtaler gjennomganger av 250 dødsfall på et sykehus i hver av de fem regioner i Danmark, altså 50 på hvert sykehus, alle typer dødsfall inkludert. Det er benyttet en tradisjonell RCRR-metode med en første screening av to sykepleiere og så en legegjennomgang der det var funnet skade. Videre vurdering av skadeklassifiserte dødsfall ble gjort i et sentralt ekspertpanel bestående av tre leger som alle var tilknyttet Patientforsikringen, og de gikk gjennom journalene i fellesskap. I tillegg ble fem spesialleger innen kardiologi, gastromedisin, lungemedisin, ortopedi og thoraxkirurgi konsultert i tre av tilfellene (20). Man konkluderte med at 8% av dødsfallene kunne vært forebygget. Alle disse rammet eldre pasienter mellom 76 og 96 år.

Det danske skjema er oversatt til norsk og introdusert som en del av det nasjonale Pasientsikkerhetsprogrammet (http://www.pasientsikkerhetskampanjen.no/no/I+trygge+hender/Innsatsomr%C3%A5der/Gjennomgang+av+50+siste+d%C3%B8dsfall.2468.cms), og det har vært i bruk ved et par sykehus i Helse Sør Øst, bl.a. Innlandet (http://www.pasientsikkerhetsprogrammet.no/no/I+trygge+hender/Innsatsomr%C3%A5der/_attachment/2897?_ts=1462845758d). Det er ikke publisert fagfellevurderte resultater fra noen av disse gjennomgangene, som i hovedsak brukes i kvalitetssikringsøyemed, og det er derfor heller ikke kjent om det foreligger undersøkelser av validitet av dette skjemaet. Det består av to A4-sider, og hver journalgjennomgang må kunne forventes å kreve mindre tid enn metoden fra England (13).

De to omtalte metoder vi ønsker å utforske (13, 20) brukes til gjennomgang av dødsfall fra en definert periode og av klinikere utenfra, altså som **måle- og rapporteringsmetoder** ment som grunnlag for læring og forbedringsarbeid. I skjemaet fra England, PRISM2, brukes uttrykk som 'problemer i helsehjelp' og 'unngåelig', mens det i 50 siste dødsfall-skjemaet benyttes 'uønskede hendelser' og 'forebyggbart'. **Disse begrepene anses å overlappe nok i innhold til at man kan tillate seg å sammenligne de to skjemaene for bruk i et norsk sykehus**.

Melding av dødsfall med avvik i Achilles sett i lys av dødsfall vurdert å være knyttet til uønskede hendelser ved RCRR

Avviksregistere har vært en del av norsk sykehusvirkelighet i flere tiår, og intensjonen er at alle avvik de ansatte opplever i arbeidssituasjonen, ikke bare de som er knyttet til pasientbehandling, skal registreres, behandles i lederlinjen og rapporteres videre etter gitte retningslinjer. I Oslo universitetssykehus (OUS) er intensjonen at alle uventede dødsfall og uønskede hendelser som fører til eller kunne ført til dødsfall eller skade, rapporteres i det elektroniske avvikssystemet, Achilles. I 2014 ble det registrert ? hendelser i Achilles med død som konsekvens, og de aller fleste av disse ?? ble rubrisert som 'ikke forebyggbar'. Samme år var det, som nevnt over, i alt 1081 pasienter som døde under opphold i OUS.

Sari et al (21) plukket ut 1006 opphold i et stort NHS-sykehus i England og gjennomførte en RCRR for å identifisere uønskede hendelser (patient safety incidents). For de samme innleggelsene ble det undersøkt om det hadde vært rapportering av avvik i sykehusets avvikssystem, og de fant at bare 10% av alle hendelser ble identifisert ved begge metoder. 83% ble bare identifisert ved RCRR, mens 7% bare ble identifisert i avvikssystemet. En amerikansk studie av metoder for å identifisere uønskede hendelser i fire sykehus i tre forskjellige regioner sammenlignet GTT, avvikssystemrapportering og bruk av pasientsikkerhetsindikatorer (PSI) slik de fremkom i sykehusets elektroniske diagnosebaserte rapportsystem (22). Også denne konkluderte at de forskjellige systemer identifiserte ulike avvik og derfor burde supplere hverandre. Tilsvarende studier for sammenligning av metoder for å identifisere uønskede hendelser som spesifikt fører til død i sykehus er, etter det jeg har kunnet finne, så langt ikke publisert. Uansett er det av interesse for OUS å vite i hvilken grad dødsfall der det har vært uønskede hendelser under sykehusopphohldet i forkant, faktisk rapporteres i det frivillige, elektroniske avvikssystemet.

Mål med studien

1. Å validere to kjente undersøkelsesinstrumenter for vurdering av alle dødsfall i et norsk sykehus

2. Å sammenligne de to metodene med tanke på bruk i vurdering av alle dødsfall i OUS i 2014

3. Å tallfeste hvor stor andel av sykehusdødsfall i OUS i 2014 som skyldtes problemer i helsehjelp / uønskede hendelser under sykehusoppholdet (med det instrument som finnes best egnet)

4. Å vurdere hvor mange av disse problemene i helsehjelp som kunne vært unngått / forebygget

5. Å beregne hvor stor andel av dødsfall i 2014 knyttet til problemer i helsehjelp bedømt ved RCRR som var meldt i det elektroniske avvikssystemet, og hvor stor andel av de avviksmeldte dødsfall som også ble plukket opp ved journalgjennomgangen av samtlige 1081 utskrevet som død

Metoder

Det var i alt 1081 pasienter som døde i Oslo universitetssykehus i 2014. Alle disse dødsfall skal gjennomgås med tanke på om det har vært problemer i helsehjelp under sykehusoppholdet og mulighet for at dødsfallet kunne vært unngått. Samtidig skal døde pasienters data sjekkes opp mot avvikssystemet med tanke på om de har vært rapportert dit. Første del av gjennomgangen skal samtidig benyttes til å avklare hvilket av de to valgte instrumenter som er best egnet til oppgaven; PRISM2 eller 50 siste dødsfall-tilnærmingen.

PRISM2

Den adapterte RCRR-tilnærming brukt i en studie av 1000 dødsfall i ti engelske sykehus i 2009 (13) inkluderte ikke psykiatriske, obstetriske eller pediatriske pasienter, liksom heller ikke pasienter definert til å være under palliativ, terminal pleie. Screeningfasen fra den opprinnelige Harvard-metoden (3, 4) ved erfaren sykepleier er fjernet, i og med alle utvalgte dødsfall vurderes. Journalgjennomgangen foretas kun av erfaren, ekstern lege som skal vurdere om det har vært problemer i helsehjelpen knyttet til dødsfallet og om disse kunne vært unngått. Skjemaet fra 2009-studien, PRISM1, er videreutviklet med tanke på bruk av samme forskergruppe (personlig meddelelse Helen Hogan 2014), vi har fått tilgang til begge og har valgt å bruke PRISM2. Metoden innebærer en gjennomgang der det skal tas stilling til om det har vært et eller flere problemer i helsehjelpen, og deretter vurderes på en syv punkts Likert-skala om problemet kunne vært unngått. Hvert pasientdødsfall tar ca. en time å vurdere.

Det vil også etableres et panel av eksperter innen alle relevante spesialiteter som kan forespørres av de leger som gjennomgår journalene om de skulle være i tvil om dødsfallet er knyttet til svikt i helsetjenesten under det aktuelle sykehusoppholdet (14). Slik eksperthjelp noteres i skjemaet.

Det strukturerte PRISM2-skjemaet (Vedlegg 1-3) er oversatt fra engelsk til norsk etter anbefalte retningslinjer av to uavhengige oversettere med norsk som morsmål (23, 24). Deretter er skjemaet gjennomgått i møte mellom de to translatører og prosjektleder for å enes om en versjon tilpasset norsk sykehusvirksomhet. Dette skjemaet er tilbakeoversatt til engelsk av en translatør med engelsk som morsmål for å sikre validitet, og siste justering av norsk skjema ble gjort i møte mellom tilbakeoversetter og prosjektleder, der opprinnelig skjema ble sammenholdt med det tilbakeoversatte for å sikre at innholdet var det samme. Den norske versjonen av skjemaet er på plass i Achilles med tanke på elektronisk gjennomgang av alle aktuelle journaler, også obstetriske, pediatriske og psykiatriske pasienter, samt pasienter i palliativ fase.

Granskere vil gjennomgå opplæring i metoden i Achilles og får i tillegg skriftlig veiledning og støtte fra prosjektledelsen ved behov. Hele journalen, inkludert kurver og diagnostiske undersøkelser, skal gjennomgås. Demografiske og kliniske data vil inkludere alder, kjønn, avdeling for innleggelse, type innleggelse (akutt, elektiv), komorbiditet (antall tilstander) og vil i stor grad importeres direkte fra sykehusets elektroniske pasientregister. Funksjonell reduksjon basert på Charlson Index of Comorbidity (25) er inkludert i skjemaet. To erfarne leger (TSK og BB) som ikke selv har jobbet de siste år i noen av de aktuelle avdelinger der dødsfall har skjedd, gjennomgår 200 tilfeldig utvalgte dødsfall innen OUS fra 2014.

Vurderingene av de to granskere sammenlignes så for validering av inter-rater pålitelighet, i tråd med studier gjort i England, der en viss andel av dødsfall vurderes av to (26).

50 siste dødsfall

Skjemaet til Siste 50 dødsfall (Vedlegg 4) er langt mindre omfattende enn PRISM-skjemaene og har ikke vært validert tidligere. Felles tilnærming i de sykehus som har tatt metoden i bruk i Norge, synes å være at både lege og sykepleier er involvert, noen steder også stabspersoner med spesialkompetanse innen pasientsikkerhet. Det har vært legetunge panel som sammen avgjør om dødsfallet var forventet, om det skjedde en uønsket hendelse under innleggelsen og om dødsfallet potensielt var forebyggbart. I vår studie vil de en lege (TSK) og ekstern sykepleier (CRT) med betydelig erfaring i pasientsikkerhetsarbeid, gjennomgå de tilfeldig utvalgte 200 journalene alene og konkludere hver for seg, evt. supplert av innspill fra superspesialister som konsulteres (som for PRISM2). Det er av interesse å belyse forskjeller i vurderinger mellom lege og sykepleier.

Vurderingene fra de to granskere sammenlignes så for validering av inter-rater pålitelighet, på samme måte som for PRISM2-gjennomgangene.

Sammenligning av PRISM2- og 50 siste dødsfall-verktøy

En lege (TSK) vil vurdere de første 100 tilfeldig utvalgte dødsfall i PRISM2, vil så gjennomgås de neste 100 i 50 siste dødsfall og deretter vil de første 100 dødsfall vurderes i 50 siste dødsfall og de siste 100 i PRISM2, slik at alle de 200 dødsfall er vurdert av samme lege i begge systemer. Dette for å sammeligne vurderingene i de to skjemaer vedrørende uønskede hendelser / problemer i helsehjelp og forebyggbare / unngåelige dødsfall. Dersom konklusjonene ligner, kunne man tenke seg å nøye seg med å bruke det enklere skjemaet i fremtidige gjennomganger av dødsfall i sykehus med tanke på måling og rapportering.

Det skjema som finnes best egnet, vil så brukes av TSK for å fullføre vurderingen av alle dødsfall i OUS i 2014.

Melding i avvikssystem

For hver av de 1081 pasienter utskrevet som døde i 2014 skal det noteres om de har vært knyttet til uønskede hendelser meldt i Achilles under det aktuelle opphold eller i de foregående 12 måneder (14). Det forventes at de aller fleste dødsfall som ble meldt i Achilles som knyttet til avvik samme år finnes igjen blant de 1081 fra utskrivingsregisteret, men noen av de avviksmeldte dødsfall er selvmord der pasienten er eller nylig har vært i behandling. Disse vil i stor grad IKKE være registrert som "utskrevet død" og vil komme i tillegg til de 1081. Deres journaler, og evt. andre i avvikssystemet som ikke finnes igjen i 'utskrevet død', må enten legges til side, eller gjennomgås på tilsvarende måte i det valgte skjema for å vurdere evt. helsetjenestesvikt.

Statistikkyndige innen OUS deltar i forberedelsene til studiene og evalueringen av de innhentede data.

Deltagere

Prosjektansvarlig/leder: Spesialrådgiver dr. med. Trine Sand Kaastad (TSK), Stab pasientsikkerhet og kvalitet
Prosjektmedarbeidere: Seksjon for pasientsikkerhet ved Thomas Jørgensen Riiser (TJR) og spesialrådgiver Ewa Ness (EN).
Avdeling for biostatistikk, epidemiologi og helseøkonomi, Forskningsstøtte, OSS – uavklart deltaker

Førsteamanuensis PhD Christine Raaen Tvedt (CRT), Lovisenberg diakonale høgskole
Professor PhD Börje Bjelke (BB), Akershus universitetssykehus
Professor dr. med. Geir Bukholm (GB), Folkehelseinstituttet og Univ for miljø og biovitenskap
Førsteamanuensis PhD Torbjørn Wisløff (TW), Folkehelseinstituttet og UiO

Gjennomgang av 200 tilfeldig utvalgte journaler to ganger (i forskjellige verktøy) gjøres av TSK, og alle 200 gjennomgås av CRT i 50 siste dødsfall-skjemaet og av BB i PRISM2-skjemaet. Behovet for gjennomgang av de resterende 881 journaler kan vurderes på ny, i og med et representativt utvalg på 200 er trukket ut.
Statistiske aspekter ivaretas av TW og uavklart deltaker fra Forskningsstøtte
Ekspertpanelet bør ha med pasientsikkerhetskyndige klinikere med bred erfaring og farmakolog eller farmasøyt

Søknader

Da de aktuelle pasienter er døde, kan de ikke bes om skriftlig samtykke, og det må søkes om dispensasjon. Dette er helsetjenesteforskning og en kvalitetsstudie som vurderes å falle utenfor REKs mandat, og den er derfor meldt til personvernombud i OUS der tilrådning ble gitt 20. april 2015 (Vedlegg 6). Det er intensjonen å publisere resultatene av studien i tidsskrift med fagfellevurdering.

Referanser

1. Helgeland J KD, Hassani S, Lindman AS, Dimoski T, Rygh LH. 30 dagers overlevelse etter innleggelse i norske sykehus i 2010 og 2011. Nasjonalt kunnskapssenter for helsetjenesten, 2013.
2. Helgeland J KD, Hassani S, Dimoski T, Lindman AS. Overlevelse og reinnleggelser ved norske sykehus for 2012. Nasjonalt kunnskapssenter for helsetjenesten, 2013.
3. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. The New England journal of medicine. 1991;324(6):370-6.
4. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. The New England journal of medicine. 1991;324(6):377-84.
5. Localio AR, Weaver SL, Landis JR, Lawthers AG, Brenhan TA, Hebert L, et al. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. Annals of internal medicine. 1996;125(6):457-64.
6. Hayward RA, McMahon LF, Jr., Bernard AM. Evaluating the care of general medicine inpatients: how good is implicit review? Annals of internal medicine. 1993;118(7):550-6.
7. Goldman RL. The reliability of peer assessments of quality of care. JAMA : the journal of the American Medical Association. 1992;267(7):958-60.
8. Lilford R, Edwards A, Girling A, Hofer T, Di Tanna GL, Petty J, et al. Inter-rater reliability of case-note audit: a systematic review. Journal of health services research & policy. 2007;12(3):173-80.
9. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2004;170(11):1678-86.
10. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Quality & safety in health care. 2009;18(4):297-302.
11. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. Journal of clinical epidemiology. 2010;63(1):94-102.
12. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Medical care. 2000;38(2):152-61.
13. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ quality & safety. 2012;21(9):737-45.
14. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. BMC health services research. 2007;7:27.
15. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA : the journal of the American Medical Association. 2001;286(4):415-20.
16. Schioler T, Lipczak H, Pedersen BL, Mogensen TS, Bech KB, Stockmarr A, et al. [Incidence of adverse events in hospitals. A retrospective study of medical records]. Ugeskrift for laeger. 2001;163(39):5370-8.
17. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 2009;21(4):285-91.
18. Improvement IfH. Move Your Dot™: Measuring, Evaluating, and Reducing Hospital Mortality Rates (Part 1). . Boston: Institute for Healthcare Improvement: Institute for Healthcare Improvement, 2003.
19. Whittington J ST, Jacobsen D Reducing Hospital Mortality Rates (Part 2). Cambridge, MA: Institute for Healthcare Improvement, 2005.
20. COWI. Forekomst af forebyggelige dødsfald på fem danske sygehuse Dansk selskab for patientsikkerhed 2013 5. APRIL 2013. Report No.

21. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ (Clinical research ed). 2007;334(7584):79.
22. Naessens JM, Campbell CR, Huddleston JM, Berg BP, Lefante JJ, Williams AR, et al. A comparison of hospital adverse events identified by three widely used detection methods. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 2009;21(4):301-7.
23. Wild D, Eremenco S, Mear I, Martin M, Houchin C, Gawlicki M, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2009;12(4):430-40.
24. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2005;8(2):94-104.
25. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987;40(5):373-83.
26. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ (Clinical research ed). 2015;351:h3239.

#	Original English version	Omforent norsk versjon tilpasset norsk sykehusvirkelighet	English version back from Norwegian
	ADMINISTRATIVE INFORMATION TO COMPLETE FOR ALL REVIEWED DEATHS	**ADMINISTRATIVE OPPLYSNINGER SOM SKAL FYLLES INN FOR ALLE GJENNOMGÅTTE OG VURDERTE DØDSFALL**	**ADMINISTRATIVE INFORMATION THAT MUST BE FILLED IN FOR ALL REVIEWED AND ASSESSED DEATHS**
1	Age at death (years)	Alder ved død (år)	Age at death (years)
2	Sex M/F	Kjønn M/K	Gender M/W
3	Length of admission (no of days)	Innleggelsens varighet (antall dager)	Duration of admission (number of days)
4	Month of admission	Innleggelsesmåned	Month of admission
5	Day of admission (Monday to Sunday)	Innleggelsesdag (mandag til søndag)	Day of admission (Monday to Sunday)
6	Time of patient's first arrival at hospital (A&E or elsewhere). Please circle. Day (08:00-16:59) Evening (17:00-21:59) Night (22:00-07:59)	Tidspunkt da pasienten først ankom til sykehuset (legevakt, akuttmottak eller annet sted). Sett ring rundt rett alternativ. Dag (08:00-16:59) Kveld (17:00-21:59) Natt (22:00-07:59)	Time when the patient first arrived at the hospital (accident and emergency unit, emergency treatment centre or other place). Circle the correct alternative. Day (8:00 a.m. - 4:59 p.m.) Evening (5:00 p.m. - 9:59 p.m.) Night (10:00 p.m. - 7:59 a.m.)
7	How many inpatient wards/units was the patient on during this admission?	Hvor mange (ikke-polikliniske) sengeposter og enheter var pasienten innom i løpet av denne innleggelsen?	How many (non-outpatient-clinic) wards and units did the patient visit during this admission?
8	Where was the patient admitted from? Please circle. His or her own home A nursing or residential care home A hospital in another NHS trust Other (specify)	Hvor ble pasienten innlagt fra? Sett ring rundt rett alternativ. Sitt eget hjem Et pleiehjem eller en omsorgsbolig Et sykehus i et annet helseforetak Annet (spesifisér)	From where was the patient admitted? Circle the correct alternative. His/her own home A nursing home or an assisted living facility A hospital in another health trust Other (specify)
9	Type of admission. Please circle Emergency Planned/elective Other (specify)	Type innleggelse. Sett ring rundt rett alternativ. Akutt Planlagt/elektiv Annet (spesifisér)	Type of admission. Circle the correct alternative. Acute Planned/elective Other (specify)

1

A	Risk Factors We ask for the following information on all patients who have died. This allows analysis of whether some groups of patients, or some types of wards and units, are disproportionately affected by potentially avoidable deaths, so improvement efforts can be focused there.	Risikofaktorer Vi ber om følgende opplysninger om alle pasienter som har dødd. Dette muliggjør analyse av hvorvidt noen pasientgrupper eller noen typer avdelinger og enheter er uforholdsmessig berørt av potensielt unngåelige dødsfall, slik at forbedringstiltak kan konsentreres der.	Risk factors We request the following information about all patients who have died. This facilitates analysis of whether any groups of patients or any types of departments and units are disproportionately affected by potentially avoidable deaths so that improvement measures can be concentrated there.
1	Did the patient have confusion/memory problems at any point in their hospital stay? Please circle. No Yes	Hadde pasienten forvirrings-/hukommelsesproblemer på noe tidspunkt under sykehusoppholdet? Sett ring rundt rett alternativ. Nei Ja	Did the patient have problems with confusion/memory at any time during the hospitalisation? Circle the correct alternative. No Yes
2	If yes, was a diagnosis of the confusion/memory problems established? Please circle. No diagnosis of type of confusion/memory problems apparent Dementia alone Delirium alone Delirium superimposed on dementia Other type of confusion/memory problems please specify	Hvis ja: Ble det fastsatt en diagnose på forvirrings-/hukommelsesproblemene? Sett ring rundt rett alternativ. Ingen diagnose av type forvirrings-/hukommelsesproblemer fremkom Kun demens Kun delirium Delirium i forbindelse med demens Annen type forvirrings-/hukommelsesproblemer; spesifisér.	If yes: Was a diagnosis specified for the confusion/memory problems? Circle the correct alternative. There was no diagnosis of the type confusion/memory problems Only dementia Only delirium Delirium in connection with dementia Other type of confusion/memory problems; specify.
3	Did the patient have a significant mental illness (other than confusion/memory problems options above)? Please circle. No indications of a significant mental illness Clear indications of a significant mental illness Some indications of a significant mental illness but records unclear	Hadde pasienten betydelig psykisk sykdom (annet enn valgmulighetene ovenfor for forvirrings-/hukommelsesproblemer)? Sett ring rundt rett alternativ. Ingen indikasjoner på betydelig psykisk sykdom Klare indikasjoner på betydelig psykisk sykdom Noen indikasjoner på betydelig psykisk sykdom, men journalen er uklar	Did the patient have significant mental illness (other than the options above for confusion/memory problems)? Circle the correct alternative. No indications of significant mental illness Clear indications of significant mental illness Some indications of significant mental illness, but the case notes are unclear
4	Did the patient have a learning disability? Please circle. No indications of a learning disability	Hadde pasienten lærevansker? Sett ring rundt rett alternativ. Ingen indikasjoner på lærevansker	Did the patient have learning difficulties? Circle the correct alternative. No indications of learning difficulties

2

	Clear indications of a learning disability Some indications of a learning disability but records unclear	Klare indikasjoner på lærevansker Noen indikasjoner på lærevansker, men journalen er uklar	Clear indications of learning difficulties Some indications of learning difficulties, but the case notes are unclear
5	Did the patient have any of the following comorbidities? This list is based on the Charlson Index of Comorbidity. Other comorbidities can be entered in the last box.	Hadde pasienten noen av de følgende komorbiditeter? Denne listen er basert på Charlsons komorbiditetsindeks. Andre samtidige lidelser kan føres inn i den siste rubrikken.	Did the patient have any of the following co-morbidities? This list is based on Charlson's co-morbidity index. Other co-morbid disorders may be entered in the last space.
	Comorbidity Yes No	**Komorbiditet Ja Nei**	**Co-morbidity Yes No**
	Myocardial infarct	Hjerteinfarkt	Heart attack
	Congestive heart failure	Hjertesvikt	Heart failure
	Peripheral vascular disease	Perifer karsykdom	Peripheral vascular disease
	Cerebrovascular disease	Cerebrovaskulær sykdom	Cerebral vascular disease
	Hemiplegia	Hemiplegi	Hemiplegia
	Chronic lung disease	Kronisk lungesykdom	Chronic respiratory disease
	Connective tissue disease	Bindevevssykdom	Collagen disease
	Diabetes without end organ damage	Diabetes uten organskade	Diabetes without organ damage
	Diabetes with end organ damage	Diabetes med organskade	Diabetes with organ damage
	Ulcer	Magesår	Gastric ulcer
	Chronic liver disease	Kronisk leversykdom	Chronic liver disease
	Moderate or severe liver disease	Moderat eller alvorlig leversykdom	Moderate or severe liver disease
	Moderate or severe kidney disease	Moderat eller alvorlig nyresykdom	Moderate or severe renal disease
	Lymphoma	Lymfom	Lymphoma
	Leukemia	Leukemi	Leukaemia
	Non-malignant tumor	Godartet svulst	Benign tumour
	Malignant tumor	Ondartet svulst	Malignant tumour
	Metastasis	Metastase	Metastasis
	AIDS	AIDS	AIDS
	Other	Annet	Other
6	Patient condition immediately prior to the illness that led to this admission. Please circle. Fully independent	Pasientens tilstand umiddelbart før sykdommen som førte til denne innleggelsen. Sett ring rundt rett alternativ. Helt selvhjulpen	The patient's condition just before the disease that resulted in this admission. Circle the correct alternative. Completely self-reliant

	Independent in personal care, but needing help with other activities of daily living Dependant on others for personal care (washing, dressing, eating, etc.) Unable to determine; no relevant information in notes (direct or implied)	Selvhjulpen for personlig stell, men hjelpetrengende for andre gjøremål i dagliglivet Avhengig av andre for personlig stell (vask, påkledning, matinntak mm.) Ikke i stand til å trekke slutning; ingen relevant informasjon i notater (direkte eller underforstått)	Self-reliant for personal care, but dependent on assistance for other tasks in daily life Dependent on others for personal care (washing, dressing, food intake, etc.) Not able to draw a conclusion; no relevant information in notes (direct or implicit)
7	Was the patient initially assessed in A&E and/or any other short term emergency admission assessment unit? (e.g. Clinical Decision Unit, Medical or Surgical Assessment Unit, etc.). Please circle. Yes No Unable to determine	Ble pasienten først vurdert i legevakt, akuttmottak og/eller annen korttidsenhet for vurdering om akutt innleggelse? (F.eks. observasjonspost) Sett ring rundt rett alternativ. Ja Nei Ikke i stand til å trekke slutning	Was the patient first assessed in an accident and emergency unit and/or other short-term unit for assessment with regard to acute admission? (E.g. observation unit) Circle the correct alternative. Yes No Not able to draw a conclusion
8	Speciality at time of first ward admission (the first ward/unit where the intention was for them to stay at least one night). Please circle. Older people's Medicine Rehabilitation General medicine (including medical assessment/short stay) Medical sub-specialities. Specify if can be determined: General Surgery (including surgical assessment/short stay) Surgical sub-specialities including gynaecology & orthopaedics. Specify if can be determined. Other. Specify	Spesialitet ved første avdelingsinnleggelse (den første avdelingen/enheten der hensikten var at pasienten skulle oppholde seg minst én natt). Sett ring rundt rett alternativ. Geriatri Rehabilitering Generell indremedisin (inkludert medisinsk vurdering/korttidsopphold) Medisinske subspesialiteter. Spesifisér hvis mulig å fastslå: Generell kirurgi / gastrokirurgi (inkludert kirurgisk vurdering/korttidsopphold) Andre kirurgiske spesialiteter, for eksempel gynekologi og ortopedi, samt kirurgiske subspesialiteter. Spesifisér hvis mulig å fastslå. Annet. Spesifisér	Specialisation at first department admission (the first department/unit where the intention was that the patient should be hospitalised for at least one night). Circle the correct alternative. Geriatrics Rehabilitation General internal medicine (including medical assessment/short-term hospitalisation) Medical sub-specialisations Specify if possible to determine: General surgery / gastrointestinal surgery (including surgical assessment/short-term hospitalisation) Other surgical specialisations, e.g. gynaecology and orthopaedics, plus surgical sub-specialisations. Specify if possible to determine. Other. Specify
9	Was this an appropriate type of ward for their condition?	Var dette en hensiktsmessig avdeling for pasientens tilstand?	Was this an appropriate department for the patient's

	Please circle. Yes, definitely appropriate Probably appropriate No Unable to determine	Sett ring rundt rett alternativ. Ja, definitivt hensiktsmessig Antagelig hensiktsmessig Nei Ikke i stand til å trekke slutning	condition? Circle the correct alternative. Yes, definitely appropriate Probably appropriate No Not able to draw a conclusion
10	Speciality at time of death. Please circle. Older people's Medicine Rehabilitation General medicine (including medical assessment/short stay) Medical sub-specialities. Specify if can be determined: General Surgery (including surgical assessment/short stay) Surgical sub-specialities including gynaecology & orthopaedics. Specify if can be determined. Other. Specify	Spesialitet ved dødstidspunkt. Sett ring rundt rett alternativ. Geriatri Rehabilitering Generell indremedisin (inkludert medisinsk vurdering/korttidsopphold) Medisinske subspesialiteter. Spesifisér hvis mulig å fastslå: Generell kirurgi / gastrokirurgi (inkludert kirurgisk vurdering/korttidsopphold) Andre kirurgiske spesialiteter, for eksempel gynekologi og ortopedi, samt kirurgiske subspesialiteter. Spesifisér hvis mulig å fastslå. Annet. Spesifisér	Specialisation at time of death. Circle the correct alternative. Geriatrics Rehabilitation General internal medicine (including medical assessment/short-term hospitalisation) Medical sub-specialisations Specify if possible to determine: General surgery / gastrointestinal surgery (including surgical assessment/short-term hospitalisation) Other surgical specialisations, e.g. gynaecology and orthopaedics, plus surgical sub-specialisations. Specify if possible to determine. Other. Specify
11	Was this an appropriate type of ward for their condition? Please circle. Yes, definitely appropriate Probably appropriate No Unable to determine	Var dette en hensiktsmessig avdeling for pasientens tilstand? Sett ring rundt rett alternativ. Ja, definitivt hensiktsmessig Antagelig hensiktsmessig Nei Ikke i stand til å trekke slutning	Was this an appropriate department for the patient's condition? Circle the correct alternative. Yes, definitely appropriate Probably appropriate No Not able to draw a conclusion
12	Apparent main diagnosis on admission: Note you should record the patient's **apparent main diagnosis at the point their initial medical assessment/clerking was**	Hoveddiagnose, slik den fremsto ved innleggelse: NB! Her skal du registrere pasientens **hoveddiagnose slik den fremsto på det tidspunktet da innledende medisinsk vurdering / inntaksregistrering var fullført** (du vil senere i	Main diagnosis, as it appeared at the time of admission: Note! Here you shall register the patient's **main diagnosis as it appeared at the time when the introductory medical assessment / admission registration was completed** (later in

	completed (you will have an opportunity later in the form to note if you consider this diagnosis was incorrect). Please circle.	skjemaet få anledning til å oppgi om du mener denne diagnosen var uriktig). Sett ring rundt rett alternativ.	the form you will have an opportunity to state whether you think this diagnosis was incorrect). Circle the correct alternative.
	Trauma-related diagnoses	Traumerelaterte diagnoser	Trauma-related diagnoses
	Fractured hip	Hoftebrudd	Hip fracture
	Any other falls-related diagnosis	Eventuelle andre fallrelaterte diagnoser	Other possible fall-related diagnoses
	Trauma from other cause (not fall)	Traume med annen årsak (ikke fall)	Trauma with another cause (not a fall)
	Cardiovascular diagnoses	Kardiovaskulære diagnoser	Cardiovascular diagnoses
	Stroke	Slag	Stroke
	Acute coronary syndrome/STEMI/angina syndrome/STEMI/angina	Akutt koronarsyndrom / STEMI / angina	Acute coronary syndrome / STEMI / angina
	Heart failure	Hjertesvikt	Heart failure
	Arrhythmia	Arytmi	Arrhythmia
	DVT/PE	Dyp venetrombose / lungeemboli	Deep vein thrombosis / pulmonary embolism
	Any other cardiovascular condition	Eventuell annen kardiovaskulær tilstand	Any other cardiovascular condition
	Infection (with or without sepsis)	Infeksjon (med eller uten sepsis)	Infection (with or without sepsis)
	Chest infection/pneumonia	Luftveisinfeksjon / lungebetennelse	Respiratory tract infection / pneumonia
	Urinary tract infection	Urinveisinfeksjon	Urinary tract infection
	Bloodstream infection	Bakteriemi	Bacteraemia
	Gastroenteritis	Gastroenteritt	Gastroenteritis
	Any other diagnosis of infection	Annen infeksjonsdiagnose	Other diagnosis of infection
	Cancer-related diagnosis	Kreftrelatert diagnose	Cancer-related diagnosis
	Acute abdomen	Akutt abdomen	Acute abdomen
	Gastrointestinal haemorrhage	Gastrointestinal blødning	Gastrointestinal haemorrhage
	Exacerbation of Chronic Obstructive Pulmonary Disease	Forverring av kronisk obstruktiv lungesykdom	Deterioration of chronic obstructive pulmonary disease
	Any other diagnosis please specify……..	Annen diagnose, spesifisér……	Other diagnosis, specify……
	We recognise the list above is not comprehensive, but it represents the diagnoses most commonly seen in patients who died in hospital in the PRISM 1 study, and will be built on in future phases.	Vi erkjenner at listen ovenfor ikke er uttømmende, men den representerer de diagnosene som var mest utbredt blant pasienter som døde på sykehus i England (PRISM1-studien)	We acknowledge that the list above is not exhaustive, but it represents the diagnoses that were most extensive among patients who died at hospitals in England (the PRISM1 study)
B	**DECISION TO PROCEED TO DETAILED REVIEW**	**BESLUTNING OM Å GÅ TIL DETALJERT**	**DECISION TO GO TO A DETAILED REVIEW AND**

	Clinical reviewer completes for all reviewed deaths Before answering the following questions, ensure you have reviewed all available documentation related to the admission in which the patient died, including: all inpatient documentation related to the admission in which the patient died, including medical, nursing and therapy records any GP referral letters, ambulance summary, A&E summary, etc. related to the admission in which the patient died **Determination of problem in healthcare**	**GJENNOMGANG OG VURDERING** **Fylles ut for alle gjenomgåtte dødsfall** Før du svarer på disse spørsmålene, må du forsikre deg om at du har gjennomgått og vurdert all tilgjengelig dokumentasjon knyttet til innleggelsen da pasienten døde, inkludert: all pasientdokumentasjon fra sykehuset knyttet til den innleggelsen der pasienten døde, inkludert medisinske, pleierelaterte og andre helserelaterte journalnotater eventuelle henvisningsbrev fra allmennlege, ambulansenotater, notater fra legevakt osv. knyttet til den innleggelsen da pasienten døde **Påvisning av problem i helsehjelp**	**ASSESSMENT** **To be filled out for all reviewed deaths** Before you answer these questions, you must assure yourself that you have reviewed and assessed all available documentation related to the admission when the patient died, including: all patient documentation from the hospital related to the admission where the patient died, including medical, care-related and other health-related case notes any referral letters from GP, ambulance notes, notes from the accident and emergency unit, etc. related to the admission when the patient died **Detection of problems in health care**
13	**A problem in healthcare is defined as 'any point where the patient's healthcare fell below an acceptable standard and led to harm'. Considering all that you know about this patient's admission, were there any problems in healthcare (including any problems before admission)? Please circle.** No evidence of any problems in healthcare ⇨ please go straight to Part D Some evidence of problem/s in healthcare ⇨ please complete the next question	**Et helsehjelpsproblem defineres som "et hvilket som helst tidspunkt der helsehjelpen til pasienten var under en akseptabel standard og førte til skade". Tatt i betraktning alt du vet om denne pasientens innleggelse: Var det noen problemer med helsehjelpen (inkludert eventuelle problemer før innleggelse)? Sett ring rundt rett alternativ.** Intet belegg for noe helsehjelpsproblem ⇨ vennligst gå rett til Del D Et visst belegg for helsehjelpsproblem(er) ⇨ svar på neste spørsmål	**A health-care problem is defined as "any point in time when the health care to the patient was below an acceptable standard and resulted in injury". Taking into consideration everything you know about this patient's admission: Were there any problems with the health care (including any problems prior to admission)? Circle the correct alternative.** No indication of any health-care problems ⇨ please go directly to Part D Some indication of health-care problem(s) ⇨ answer the next question
14	**In your judgement, is there some evidence that the patient's death was avoidable if the problem/s in healthcare had not occurred? Please circle.** No , death was definitely not avoidable ⇨ please go straight to Part D At least slight evidence the death may have been avoidable ⇨ please complete Part C and then Part D	**Slik du vurderer det; finnes det noe belegg for at pasientens død kunne ha vært unngått dersom helsehjelpsproblemet/-ene ikke hadde forekommet? Sett ring rundt rett alternativ.** Nei, dødsfallet var definitivt ikke mulig å unngå ⇨ gå rett til Del D I det minste et snev av belegg for at dødsfallet kunne ha vært mulig å unngå ⇨ fyll ut Del C og så Del D	**The way you assess it; is there any indication that the patient's death could have been avoided if the health-care problem(s) had not occurred? Circle the correct alternative.** No, it was definitely not possible to avoid the death ⇨ go directly to Part D At least a slight indication that it could have been possible to avoid the death ⇨ fill out Part C and then Part D

C	**DETAILED REVIEW OF PROBLEMS IN HEALTHCARE** **Clinical reviewers complete this section ONLY if you have answered Question 14 as "At least slight evidence the death may have been avoidable."** **Please summarise in chronological order the background, admission, procedures, and events leading up to the patient's death and cause of death, including any points where there were problems in healthcare. You will have an opportunity to be more specific about these problems in healthcare and justify your judgements later in the review form.**	**DETALJERT GJENNOMGANG OG VURDERING AV PROBLEMER I HELSEHJELP** **Dette avsnittet fylles KUN ut dersom du har besvart spørsmål 14 med "I det minste et snev av belegg for at dødsfallet kunne ha vært mulig å unngå."** **Oppsummér kronologisk bakgrunnen, innleggelsen, prosedyrene og hendelsene som ledet frem til pasientens død, samt dødsårsaken, inkludert eventuelle punkter der det var helsehjelpsproblemer. Du vil få mulighet til å gå nærmere inn på disse helsehjelpsproblemene og begrunne dine bedømmelser senere i vurderingsskjemaet.**	**DETAILED REVIEW AND ASSESSMENT OF PROBLEMS IN HEALTH CARE** **This section should ONLY be filled out if you have answered question 14 with "At least a slight indication that it could have been possible to avoid the death."** **Summarise chronologically the background, the admission, the procedures and the events that led to the patient's death, together with the cause of death, including any instances where there were health-care problems. You will have an opportunity to go into more detail about these health-care problems and justify your appraisals later in the assessment form.**
15	**Please complete the following table using the category list that accompanies this table which is at the end of the review form.** **- A problem in healthcare is defined as 'any point where the patient's healthcare fell below an acceptable standard and led to harm'.** To identify the **problems in healthcare,** consider what an acceptable standard of healthcare would be for this patient, and articulate how the healthcare they received fell below this acceptable standard (whether through omission, delay or incorrect actions). Include any problems in healthcare that occurred before the patient's final admission but were identified during it. **Only one problem should be entered per row.** - It can be difficult to identify **contributory factors** (i.e. the underlying reasons ***why*** the problem in healthcare occurred) from case notes alone. If you can clearly identify any factors that contributed to each problem in healthcare please do so, but avoid making assumptions. **Contributory factor should refer to the problem described in the same row.**	**Fyll ut følgende tabell ved hjelp av den kategorilisten som ledsager tabellen og som finnes på slutten av vurdeingsskjemaet.** **Et helsehjelpsproblem defineres som "et hvilket som helst sted der helsehjelpen til pasienten var under en akseptabel standard og førte til skade".** For å identifisere **helsehjelpsproblemer**, vurder hva som ville vært en akseptabel standard for helsehjelp for denne pasienten, og uttrykk hvordan den helsehjelpen pasienten mottok ikke nådde opp til denne akseptable standarden (enten det skyldtes unnlatelse, forsinkelse eller uriktige handlinger). Ta med eventuelle helsehjelpsproblemer som forekom før pasientens siste innleggelse, men som ble identifisert i løpet av den. **Kun ett problem skal føres inn per rad.** - Det kan være vanskelig å identifisere **medvirkende faktorer** (dvs. de underliggende **årsakene** til at helsehjelpsproblemet forekom) kun ut ifra journalnotater. Hvis du klart kan identifisere faktorer som bidro til hvert av problemene i helsehjelpen, ber vi deg om å gjøre det, men unngå antagelser. **Medvirkende faktor skal henvise til problemet som beskrives i samme rad.**	**Fill out the following table with the aid of the category list that accompanies the table and that is found at the end of the assessment form.** **A health-care problem is defined as "any place where the health care to the patient was below an acceptable standard and resulted in injury".** In order to identify **health-care problems**, assess what would be an acceptable standard for health care for this patient, and state how the health care that the patient received did not meet this acceptable standard (whether it was due to neglect, delay or incorrect actions). Include any health-care problems that occurred before the patient's last admission, but that were identified during that admission. **Only one problem should be entered per row.** - It may be difficult to identify **contributing factors** (i.e. the underlying **reasons** why the health-care problem occurred) on the basis of the case notes alone. If you can clearly identify factors that contributed to each of the problems with health care, we request that you do so, but avoid assumptions. **A contributing factor shall refer to the problem that is described on the same row.**
Ta	**Describe each problem in care in your own words.**	**Beskriv hvert helsehjelpsproblem med dine egne ord.**	**Describe each health-care problem in your own words.**

b 1	**Please articulate what should have happened AND what did happen.**	**Beskriv hva som burde ha skjedd OG hva som skjedde.**	**Describe what ought to have happened AND what did happen.**
2	**Where did the problem occur?**	**Hvor forekom problemet?**	**Where did the problem occur?**
3	**Sub-type of problem** ***(select one)***	**Undertype av problem** ***(velg én)***	**Sub-type of problem** ***(choose one)***
4	**Contributory factors** ***(option to select none, one or multiple)***	**Medvirkende faktorer** ***(du kan velge ingen, én eller flere)***	**Contributing factors** ***(you may choose none, one or several)***
Ek s1	*Example: "First dose of IV penicillin should have been given immediately but was not given until three hours after prescribed*	Eksempel: "Første dose av IV penicillin burde ha blitt gitt umiddelbart, men ble først gitt tre timer etter at det var forskrevet	Example: "First dose of IV penicillin ought to have been given immediately, but was first given three hours after it was prescribed
16	**Earlier in the case note review, you made a judgement that there was at least slight evidence that death may have been avoidable if the problem/s in healthcare had not occurred. Considering the problems in healthcare you have described above, please rate on the Likert scale the strength of evidence for the avoidability of the death:** Slight evidence for avoidability Possibly avoidable but not very likely, less than 50-50 but close call Probably avoidable, more than 50-50 but close call Strong evidence for avoidability Definitely avoidable	**Tidligere i gjennomgangen av journalen vurderte du det slik at det i det minste var et snev av belegg for at dødsfallet kunne ha vært mulig å unngå hvis problemet/problemene i helsehjelpen ikke hadde funnet sted. I lys av de problemene i helsehjelpen du har beskrevet over, angi på Likert-skalaen hvor sterkt belegg det er for at dødsfallet kunne ha vært unngått:** Snev av belegg for at det kunne ha vært unngått Mulig unngåelig, men ikke veldig sannsynlig; mindre enn 50-50, men i nærheten Sannsynligvis unngåelig, mer enn 50-50, men i nærheten Sterkt belegg for at det kunne ha vært unngått Kunne definitivt ha vært unngått	**Earlier in the review of the case notes, you assessed it to be the case that there was at least a slight indication that it could have been possible to avoid the death if the problem(s) with health care had not occurred. In light of the problems with health care that you have described above, indicate on the Likert scale how strong the indication is that the death could have been avoided:** Slight indication that it could have been avoided Possibly avoidable, but not very likely; less than 50-50, but close to that Probably avoidable, more than 50-50, but close to that Strong indication that it could have been avoided Could definitely have been avoided
	Please record reasons justifying the judgement you have made	**Angi begrunnelsene for den vurderingen du har gjort**	**State the grounds for the assessment you have made**
17	**Earlier in the case note review, you made a judgement that there was at least slight evidence that death may have been avoidable if the problem/s in healthcare had not occurred. Considering the problems in healthcare you have described above, please mark on this continuous scale the strength of evidence for the avoidability of the death. <u>Mark with a single line through the scale.</u>** We appreciate this is an even more difficult judgement call than the decision you made above on Likert Scale (slight/possible/probable/strong evidence for avoidability), but providing your judgement on a continuous scale allows	**Tidligere i journalgjennomgangen vurderte du det slik at det i det minste fantes et snev av belegg for at dødsfallet kunne ha vært mulig å unngå dersom problemet/-ene i helsehjelpen ikke hadde forekommet. I lys av de problemene i helsehjelpen du har beskrevet over, vennligst angi på denne kontinuerlige skalaen hvor sterkt belegget er for at dødsfallet kunne ha vært unngått. <u>Merk med én enkelt strek gjennom skalaen.</u>** Vi forstår at dette er en enda vanskeligere skjønnsmessig vurdering å foreta enn den vurderingen du gjorde ovenfor på	**Earlier in the review of the case notes, you assessed it to be the case that there was at least a slight indication that it could have been possible to avoid the death if the problem(s) with health care had not occurred. In light of the problems with health care that you have described above, please indicate on this continuous scale how strong the indication is that the death could have been avoided: <u>Mark with a single line through the scale.</u>** We understand that this is an even more difficult discretionary

	additional analysis.	Likert-skalaen (et snev av / mulig / sannsynlig / sterkt belegg for at dødsfallet kunne vært unngått), men det at du angir din vurdering på en glidende skala, muliggjør ytterligere analyse.	assessment to undertake than the assessment you made above on the Likert scale (a slight / possible / probable / strong indication that the death could have been avoided), but indicating your assessment on a sliding scale makes further analysis possible.
	Death definitely not avoidable	Dødsfallet definitivt ikke unngåelig	The death definitely Unavoidable
	Death definitely avoidable	Dødsfallet definitivt unngåelig	The death definitely Avoidable
18	**If death was considered avoidable had the problem/s in healthcare had not occurred, by how many days/months/years do you estimate the patient's life was shortened? Please circle.** By one week or less By more than a week but less than a month By more than a month but less than three months By more than three months but less than a year By ……years We appreciate this is difficult judgement call, but even estimates are helpful in prioritising future improvement efforts. In arriving at an estimate, you may wish to consider expected prognosis for a patient presenting with this condition and comorbidities who received an acceptable standard of healthcare, and/or average life expectancy alongside consideration of whether the patient had better or worse general health and capacity to recover than average.	**Hvis du anser at dødsfallet kunne ha vært unngått dersom problemet/-ene i helsehjelpen ikke hadde forekommet: Med hvor mange dager/måneder/år anslår du at pasientens liv ble forkortet? Sett ring rundt rett alternativ.** Med én uke eller mindre Med mer enn en uke men mindre enn en måned Med mer enn en måned men mindre enn tre måneder Med mer enn tre måneder men mindre enn et år Med ……. år Vi forstår at dette er en vanskelig skjønnsmessig vurdering å foreta, men selv anslag er til hjelp i prioriteringen av fremtidige forbedringstiltak. Når du skal utarbeide et anslag, kan du gjerne ta utgangspunkt i forventet prognose for en pasient med denne tilstanden og disse komorbiditetene som hadde mottatt helsehjelp av akseptabel standard, og/eller gjennomsnittlig forventet levealder og hvorvidt pasienten hadde bedre eller dårligere allmenn helse og tilfriskningsevne enn gjennomsnittet.	**If you think that the death could have been avoided if the problem(s) with health care had not occurred: by how many days, months/years do you estimate that the patient's life was shortened? Circle the correct alternative.** By one week or less By more than one week but less than a month By more than a month but less than three months By more than three months but less than a year By ……. years We understand that this is a difficult discretionary assessment to make, but even an estimate is helpful in the prioritisation of future improvement measures. When you are going to make an estimate, please base it on the expected prognosis for a patient with this condition and the co-morbidities that had received health care of an acceptable standard, and/or the average life expectancy and the extent to which the patient had better or worse than average general health and capability of recovery.
19	**If death was considered avoidable had the problem/s in healthcare not occurred, please indicate when you believe the BEST opportunity of avoiding the death occurred:** Outside hospital care (primary care, ambulance, etc.) In a prior admission/attendance (this trust) In a prior admission/attendance (another secondary healthcare provider)	**Hvis du anser at dødsfallet kunne ha vært unngått dersom helseproblemet/-ene i helsehjelpen ikke hadde forekommet: Angi når du tror den BESTE anledningen til å unngå dødsfallet bød seg:** Utenfor sykehuset (primærhelsetjeneste, ambulanse, osv) Ved tidligere innleggelse/oppmøte (ved dette helseforetaket)	**If you think that the death could have been avoided if the health problem(s) with health care had not occurred: Indicate when you think the BEST opportunity for avoiding the death presented itself:** Outside the hospital (primary health service, ambulance, etc.) On a previous admission/appointment (at this health trust)

	In an initial assessment unit (e.g. A&E department, or any other short term emergency assessment unit such as a Clinical Decision Unit, Medical Assessment Unit, Surgical Assessment Unit, etc.) During an invasive procedure (including surgery and anaesthesia) During post-operative care or post-procedure care (except HDU/ITU) During High Dependency or ITU care (not including decision to refer to HDU/ITU) During inpatient care on a ward/unit designated as: Older people's Medicine Rehabilitation General medicine Medical sub-specialities General Surgery Surgical sub-specialities including gynaecology & orthopaedics Other (specify)	Ved tidligere innleggelse/oppmøte (annen leverandør av sekundære helsetjenester) Ved en enhet for prehospital vurdering (f.eks. legevakt, akuttmottak eller annen korttidsenhet for vurdering for akutt innleggelse, som en observasjonspost osv.) Under en invasiv prosedyre (inkludert kirurgi og anestesi) Under postoperativ overvåking eller overvåking etter prosedyrer (unntatt intensivavdeling) Under behandling i intensivavdeling (ikke inkludert beslutning om henvisning til intensivavdeling) Som innlagt ved avdeling/enhet for: Geriatri Rehabilitering Generell indremedisin Medisinske subspesialiteter Generell kirurgi Andre kirurgiske spesialiteter, for eksempel gynekologi og ortopedi, samt kirurgiske subspesialiteter. Annet (spesifisér)	On a previous admission/appointment (other provider of secondary health services) At a unit for pre-hospital assessment (e.g. accident and emergency unit, emergency treatment centre or other short-term unit for assessment for acute admission, such as an observation unit etc.) During an invasive procedure (including surgery and anaesthesia) During post-operative monitoring or monitoring after procedures (with the exception of the intensive care unit) During treatment in an intensive care unit (not including decisions regarding referral to an intensive care unit) When admitted to a department/unit for: Geriatrics Rehabilitation General internal medicine Medical sub-specialisations General surgery Other surgical specialisations, e.g. gynaecology and orthopaedics, plus surgical sub-specialisations. Other (specify)

20	**Avoiding future deaths** Although case note review in isolation cannot be a substitute for a full root cause analysis investigation, please indicate any specific improvements you believe might decrease the likelihood of similar deaths occurring in future. Areas you might consider are better design of equipment or procedures, interventions to limit human error or organisational changes.	**Unngåelse av fremtidige dødsfall** Selv om gjennomgang av journal isolert sett ikke kan erstatte en fullstendig undersøkelse av de underliggende årsakene, ber vi deg angi eventuelle konkrete forbedringer som du mener kan minske sannsynligheten for at lignende dødsfall vil forekomme i fremtiden. Noen områder du kanskje kan vurdere, er bedre utforming av utstyr eller prosedyrer, inngrep for å begrense menneskelige feil eller organisasjonsmessige endringer.	**Avoidance of future deaths** Although a review of the case notes in isolation cannot replace a complete examination of the underlying causes, we request that you indicate any specific improvements that you think may reduce the likelihood that similar deaths will occur in the future. Some areas that you may be able to assess are improvement of equipment or procedures, procedures for limiting human error or organisational changes.
D	**GENERAL QUALITY OF CARE AND END OF LIFE CARE** **Complete for ALL reviewed deaths** **Overall Quality of Care**	**GENERELL KVALITET PÅ HELSEHJELP OG OMSORG VED LIVETS SLUTT** **Fyll inn for ALLE gjennomgåtte og vurderte dødsfall** **Helhetlig kvalitet på helsehjelp**	**GENERAL QUALITY OF HEALTH SERVICES AND SERVICES AT THE END OF LIFE** **Fill in for ALL reviewed and assessed deaths** **Consistent quality of health care**
21	Considering all that you know about this patient's admission, how would you rate the OVERALL quality of healthcare received by the patient from this trust? This question recognises that a problem in care causing patient harm can occur against a backdrop of overall good quality care, and the converse, a patient may experience poor overall quality of care without obvious harm. For this question, do not consider healthcare prior to the admission that ended in the patient's death or give detail of a specific problem in care causing harm, which were entered in Part C. Excellent Good Adequate Poor Very poor	Med tanke på alt du vet om denne pasientens innleggelse: Hvordan vil du klassifisere den HELHETLIGE kvaliteten på helsehjelpen som pasienten fikk fra dette helseforetaket? Dette spørsmålet anerkjenner at et problem i helsehjelpen som påfører pasienten skade kan skje mot en bakgrunn av helsehjelp av generelt god kvalitet, og motsatt, at en pasient kan oppleve helsehjelp av generelt dårlig kvalitet uten noen tilsynelatende skade. I besvarelsen av dette spørsmålet skal du ikke ta hensyn til helsehjelp før innleggelsen som endte med pasientens død eller gi detaljer om et spesifikt problem i helsehjelp som gjorde skade, som ble lagt inn i Del C. Utmerket God	Keeping in mind everything you know about this patient's admission: How would you classify the CONSISTENT quality of the health care that the patient received from this health trust? This question recognizes that a problem in health care that inflicts an injury on the patient may occur with a background of health care of generally good quality, and on the contrary, that a patient may feel that health care is of a general poor quality without any apparent injury. In the answer to this question, you should not take into consideration health care prior to the admission that ended with the patient's death or give details about a specific problem in the performance that caused injury that was entered in Part C. Excellent

		Tilstrekkelig Dårlig Svært dårlig	Good Satisfactory Poor Very poor
	Please add any detail on overall quality of healthcare that can be used for learning (positive or negative):	Legg gjerne til ytterligere betraktninger om helhetlig kvalitet på helsehjelpen som kan brukes til læring (positive eller negative):	Please add further observations about the general quality of the health care that can be used for learning (positive or negative):
	End of Life Care **Questions 22 and 23 focus on care EITHER from the point where the patient was recognised at high risk of dying (whether this was days or hours before death) OR, for patients who were not recognised as at high risk of dying, the last 48 hours of their life**	**Helsehjelp ved livets slutt** **Spørsmål 22 og 23 fokuserer på helsehjelp ENTEN fra det tidspunktet da man erkjente at risikoen for at pasienten skulle dø var høy (enten dette var dager eller timer før døden inntraff) ELLER, for pasienter man ikke erkjente var i store fare for å dø, de siste 48 timene av livet.**	**Health care at the end of life** **Questions 22 and 23 focus on health care EITHER from the time when it was recognised that the risk that the patient would die was high (whether this was days or hours before death occurred) OR for patients where it was not recognised that they were in great danger of dying during the last 48 hours of their life.**
22	Was the patient subject to any intrusive or invasive procedures that were **not** in their best interests at the end of life (including inappropriate attempts at CPR)? Yes No Unable to determine	Ble pasienten utsatt for noen inngripende og invasive prosedyrer som **ikke** var til vedkommendes beste ved livets slutt (inkludert utilbørlige forsøk på hjertelungeredning)? Ja Nei Ikke i stand til å trekke slutning	Was the patient subjected to any invasive procedures that were **not** for the patient's best at the end of life (including unwarranted attempts at cardiopulmonary resuscitation)? Yes No Not able to draw a conclusion
23	Was there evidence of discussion of end of life care with family/friends? Please circle. Yes, evidence of discussion No, discussion appeared appropriate and feasible, but no evidence it took place Not appropriate/not feasible to discuss with family/friends	Fantes det belegg for at omsorg ved livets slutt hadde blitt diskutert med familie/venner? Sett ring rundt rett alternativ. Ja, belegg for diskusjon Nei, diskusjon hadde tilsynelatende vært hensiktsmessig og gjennomførbar, men det er ikke noe belegg for at det fant sted Ikke hensiktsmessig/ikke gjennomførbart å diskutere med familie/venner	Were there indications that care at the end of life had been discussed with family/friends? Circle the correct alternative. Yes, indications of discussion No, discussion had apparently been appropriate and feasible, but there is not any indication that it took place Inappropriate/infeasible to discuss with family/friends
	Please add any detail on overall quality of end of life healthcare that can be used for learning (positive or negative) including pain and symptom control:	Legg gjerne til ytterligere betraktninger om helhetlig kvalitet på helsehjelpen ved livets slutt som kan brukes til læring (positive eller negative), inkludert smerte- og symptomkontroll:	Please add further observations about the consistent quality of the health care at the end of life that can be used for learning (positive or negative), including control of pain and symptoms:
E	**REVIEW PROCESS INFORMATION** **Complete for ALL reviewed deaths**	**INFORMASJON OM GJENNOMGANGS- OG VURDERINGSPROSESSEN** **Fyll inn for ALLE gjennomgåtte dødsfall**	**INFORMATION ABOUT THE REVIEW AND ASSESSMENT PROCESS** **Fill in for ALL reviewed deaths**
24	Were your judgements limited or hampered by lack of	Ble dine vurderinger begrenset eller hemmet av mangel på	Were your assessments limited or inhibited by a lack of sub-

	subspecialty knowledge? No Yes	subspesialitetskompetanse? Nei Ja	specialist expertise? No Yes
25	If so was a second specialist opinion sought? No Yes	Hvis ja: Ble det innhentet vurdering fra en annen spesialist? Nei Ja	If yes: was an assessment obtained from another specialist? No Yes
26	What was your question/s for the specialist?	Hvilke(t) spørsmål stilte du spesialisten?	What question(s) did you ask the specialist?
27	What was the answer/s from the specialist?	Hvilke(t) svar ga spesialisten?	What answer(s) did the specialist give?
28	Did the answer/s change your opinion and how?	Endret svaret/svarene din oppfatning, og hvordan?	Did the answer(s) change you opinion, and if so, how?
29	**How adequate were the records in providing information to enable judgements of problems in care? Please circle.** Medical records were adequate to make a reasonable judgement Some deficiencies in the records (specify) Major deficiencies (specify) Severe deficiencies, impossible to make judgements about problems in care	**I hvilken grad ga journalene tilstrekkelig informasjonsgrunnlag til å bedømme helsehjelpsproblemer? Sett ring rundt rett alternativ.** Pasientjournalene var tilstrekkelige til at man kunne foreta en rimelig bedømmelse Noen mangler ved journalene (spesifisér) Betydelige mangler (spesifisér) Alvorlige mangler, umulig å bedømme mulige helsehjelpsproblemer	**To what extent did the case notes provide sufficient information to appraise health-care problems? Circle the correct alternative.** The patient records were sufficient to enable us to make a reasonable appraisal Some shortcomings in the case notes (specify) Substantial shortcomings (specify) Serious shortcomings, impossible to appraise possible health-care problems
	Please use this space to specify any deficiencies in the medical record	Bruk gjerne plassen til å spesifisere eventuelle mangler ved pasientjournalen	Please use this space to specify any shortcomings in the patient records
30	**Total time taken to complete** review (minutes)?	**Total tid brukt på** gjennomgangen (minutter)?	**Total time spent on** the review (minutes)?

#	Original English	Omforent norsk versjon med forbedringer	English
	OPTIONS FOR PROBLEMS TABLE	**TABELL FOR PROBLEMALTERNATIVER** kategorilisten	**TABLE FOR PROBLEM ALTERNATIVES** The list of categories
	Where did the problem occur?	**Hvor forekom problemet?**	**Where did the problem occur?**
	Outside hospital care (primary care, ambulance, etc.)	Helsetjenester utenfor sykehuset (primærhelsetjenester, ambulanse, osv.)	Health services outside the hospital (primary health service, ambulance, etc.)
	In a prior admission/attendance (this trust)	Ved tidligere innleggelse/oppmøte (dette helseforetaket)	On a previous admission/appointment (this health trust)
	In a prior admission/attendance (another secondary healthcare provider)	Ved tidligere innleggelse/oppmøte (annen sekundær helsetjenesteleverandør)	On a previous admission/appointment (other secondary health service provider)
	During this admission (first 48 hours)	I løpet av denne innleggelsen (første 48 timer)	During this admission (first 48 hours)
	During this admission (after the first 48 hours)	I løpet av denne innleggelsen (etter de første 48 timene)	During this admission (after the first 48 hours)
	Type of problem in healthcare	**Type helsehjelpsproblem**	**Type of health-care problem**
	Sub-type of problem in healthcare	**Undertype helsehjelpsproblem**	**Sub-type of health-care problem**
A	Problem in assessment, investigation or diagnosis *(including assessment of pressure ulcer risk, VTE risk, history of falls)*	Problem innen vurdering, utredning eller diagnose *(inkludert vurdering av risiko for trykksår, tromboserisiko, fallhistorikk)*	Problems with assessment, examination or diagnosis *(including assessment of risk of bedsores, risk of thrombosis, history of falls)*
	A1.Physical examination and history taking	A1. Fysisk undersøkelse og anamneseopptak	A1. Physical examination and recording of medical history
	A2.Pressure ulcer risk not assessed/incorrectly assessed	A2. Risiko for trykksår ikke vurdert/feilvurdert	A2. Risk of bedsores not assessed/incorrectly assessed
	A3.VTE risk assessment not completed/incorrectly completed	A3. Tromboserisikovurdering ikke gjennomført/gjennomført feil	A3. Assessment of risk of thrombosis not performed/incorrectly performed
	A4.Falls history/vulnerability to falls not identified	A4. Fallhistorikk/sårbarhet for fall ikke identifisert	A4. History of falls/vulnerability for falls not identified
	A5.Swallowing safety not assessed/incorrectly assessed	A5. Trygghet ved svelging ikke vurdert/feilvurdert	A5. Safety when swallowing not assessed/incorrectly assessed
	A6.Tests and investigations missed/delayed/wrong	A6. Undersøkelser og utredninger oversett/forsinket/feil	A6. Examinations and studies omitted/delayed/incorrect
	A7.Diagnosis missed/delayed/wrong	A7. Diagnose oversett/forsinket/feil	A7. Diagnosis overlooked/delayed/incorrect
	A.Other assessment, investigation or diagnosis	A. Annet vurderings-, utrednings- eller	A. Other problems with assessment, examination or

15

	problem	diagnoseproblem	diagnosis
B	Problem with medication/IV fluids/electrolytes/oxygen *(other than anaesthetic)*	Problem med medisinering/IV-væsker /elektrolytter/oksygen *(annet enn anestesimiddel)*	Problems with medication/IV fluids /electrolytes/oxygen *(other than anaesthetic)*
	B1.Overhydration	B1. Overhydrering	B1. Over-hydration
	B2.Underhydration	B2. Underhydrering	B2. Under-hydration
	B3.Oxygen supply wrong/delayed/omitted	B3. Oksygentilførsel feil/forsinket/utelatt	B3. Oxygen supply incorrect/delayed/omitted
	B4.Allergic/anaphylactic reaction to any medication	B4. Allergisk/anafylaktisk reaksjon på en eller annen medisinering	B4. Allergic/anaphylactic reaction to one medication or another
	B5.Anticoagulants/antiplatelets wrong/delayed/omitted	B5. Antitrombosemidler/platehemmere feil/forsinket/utelatt	B5. Antithrombotic agents/platelet inhibitors incorrect/delayed/omitted
	B6.Antibiotics wrong/delayed/omitted	B6. Antibiotika feil/forsinket/utelatt	B6. Antibiotics incorrect/delayed/omitted
	B7.Insulin or other diabetes medication wrong/delayed/omitted	B7. Insulin eller annen diabetesmedisinering feil/forsinket/utelatt	B7. Insulin or other diabetes medication incorrect/delayed/omitted
	B8.Opiates wrong/delayed/omitted	B8. Opiater feil/forsinket/utelatt	B8. Opiates incorrect/delayed/omitted
	B9.Sedatives/hypnotics/antipsychotics wrong/delayed/omitted	B9. Sedativer/hypnotika/antipsykotika feil/forsinket/utelatt	B9. Sedatives/hypnotics/anti-psychotics incorrect/delayed/omitted
	B10.Steroids wrong/delayed/omitted	B10. Steroider feil/forsinket/utelatt	B10. Steroids incorrect/delayed/omitted
	B11.NSAID wrong/delayed/omitted	B11. NSAID feil/forsinket/utelatt	B11. NSAID incorrect/delayed/omitted
	B12.Diuretics wrong/delayed/omitted	B12. Diuretika feil/forsinket/utelatt	B12. Diuretics incorrect/delayed/omitted
	B13.Antihypertensives wrong/delayed/omitted	B13. Antihypertensiva feil/forsinket/utelatt	B13. Anti-hypertensives incorrect/delayed/omitted
	B14.Cardiovascluar medications wrong/delayed/omitted	B14. Hjertekarmedisinering feil/forsinket/utelatt	B14. Cardiovascular medication incorrect/delayed/omitted
	B15.Chemotherapy wrong/delayed/omitted	B15. Kjemoterapi feil/forsinket/utelatt	B15. Chemotherapy incorrect/delayed/omitted
	B16.Other medication/IV fluids/electrolytes/oxygen problem	B16. Problem med annen medisinering/IV-væsker /elektrolytter/oksygen	B16. Problems with other medication/IV fluids /electrolytes/oxygen
C	Problem related to treatment and management plan *(including prevention of pressure ulcers, falls, VTE*	Problem knyttet til behandlings- og tiltaksplan *(inkludert forebygging av trykksår, fall, trombose)*	Problems related to treatment and action plan *(including prevention of bedsores, falls, thrombosis)*
	C1.Appropriate medical/surgical treatment not planned	C1. Hensiktsmessig medisinsk/kirurgisk behandling ikke planlagt	C1. Appropriate medical/surgical treatment not planned
	C2.Avoidable delay/omission of planned medical/surgical treatment	C2. Unngåelig forsinkelse/utelatelse av planlagt medisinsk/kirurgisk behandling	C2. Avoidable delay/omission of planned medical/surgical treatment

16

	C3.Inappropriate/unnecessary medical/surgical treatment given	C3. Uhensiktsmessig/unødvendig medisinsk/kirurgisk behandling gitt	C3. Inappropriate/unnecessary medical/surgical treatment given
	C4.Inappropriate ceiling of care	C4. Uhensiktsmessig tak for helsehjelp	C4. Inappropriate ceiling for health care
	C5.Omitted/delayed/wrong treatment from AHPs	C5. Utelatt/forsinket/feil behandling fra alternative helsetjenesteaktører	C5. Omitted/delayed/incorrect treatment from alternative health service providers
	C6.Acquired pressure ulcer: prevention below acceptable standard	C6. Oppstått trykksår: forebygging dårligere enn akseptabel standard	C6. Bedsores occurred: prevention worse than acceptable standard
	C7.Acquired pressure ulcer despite apparently acceptable standard of prevention	C7. Oppstått trykksår til tross for tilsynelatende akseptabel standard på forebygging	C7. Bedsores occurred despite apparently acceptable standard for prevention
	C8.Slip/trip/fall: prevention plan below acceptable standard	C8. Skli-/snublehendelse/fall: forebyggingsplan dårligere enn akseptabel standard	C8. Slip/stumble event/fall: prevention plan worse than acceptable standard
	C9.Slip/trip/fall despite apparently acceptable standard of falls prevention	C9. Skli-/snublehendelse/fall til tross for tilsynelatende akseptabel standard på fallforebygging	C9. Slip/stumble event/fall despite apparently acceptable standard for prevention of falls
	C10.Developed VTE: prophylaxis below acceptable standard	C10. Utviklet DVT/emboli: profylakse dårligere enn akseptabel standard	C10. Developed DVT/embolism: prophylactic worse than acceptable standard
	C11.Developed VTE despite apparently acceptable standard of VTE prophylaxis	C11. Utviklet DVT/emboli til tross for tilsynelatende akseptabel standard på tromboseprofylakse	C11. Developed DVT/embolism despite apparently acceptable standard for thromboprophylaxis
	C12.Other treatment or management related problem	C12. Annet behandlings- eller håndteringsrelatert problem	C12. Other treatment or handling-related problem
D	Problem with infection control	Problem med infeksjonskontroll	Problems with infection control
	D1.Surgical wound infection	D1. Postoperativ sårinfeksjon	D1. Post-operative wound infection
	D2.Infection from invasive procedure other than surgery	D2. Infeksjon etter invasiv prosedyre utenom kirurgi	D2. Infection after invasive procedure other than surgery
	D3.Other healthcare associated wound infection *(e.g. infected ulcer)*	D3. Annen helsehjelpsrelatert sårinfeksjon *(f.eks. infisert leggsår)*	D3. Other health-service-related wound infection *(e.g. infected venous ulcer)*
	D4.Infection from indwelling device *(catheter, central lines, etc.)*	D4. Infeksjon fra innlagt enhet *(kateter, CVK-er osv.)*	D4. Infection from inserted device *(catheter, CVCs, etc.)*
	D5.Healthcare associated clostridium difficile	D5. Helsehjelpsrelatert clostridium difficile	D5. Health-care-related clostridium difficile colitis
	D6.Healthcare associated MRSA bloodstream infection	D6. Helsehjelpsrelatert bakteriemi med MRSA	D6. Health-care-related bacteraemia with MRSA
	D7.Other bloodstream infection (not MRSA)	D7. Annen bakteriemi (ikke MRSA)	D7. Other bacteraemia (not MRSA)
	D8.Healthcare associated pneumonia/chest infection	D8. Helsetjenesterelatert	D8. Health-service-related pneumonia/respiratory

17

	(including aspiration)	lungebetennelse/luftveisinfeksjon (inkludert aspirasjon)	tract infection (including aspiration)
	D9.Healthcare associated norovirus/D&V	D9. Helsetjenesterelatert norovirus/diaré og oppkast	D9. Health-service-related norovirus/diarrhoea and vomiting
	D10.Other infection control problem	D10. Annet infeksjonskontrollproblem	D10. Other infection control problems
E	Problem related to operation/invasive procedure *(other than infection control)*	Problem knyttet til operasjon/invasiv prosedyre *(annet enn infeksjonskontroll)*	Problems related to operation/invasive procedure *(other than infection control)*
	E1.Avoidable delay in undertaking procedure	E1. Unngåelig forsinkelse i utførelse av inngrep	E1. Avoidable delay in execution of intervention
	E2.Inadequate pre-procedure assessment/preparation	E2. Utilstrekkelig vurdering/forberedelse i forkant av prosedyre	E2. Insufficient assessment/preparation prior to procedure
	E3.Anaesthetic/sedation problem including airway management	E3. Anestesi-/sedasjonsproblem inkludert sikring av luftveier	E3. Anaesthesia/sedation problems including safeguarding of respiratory tract
	E4.Problem related to operative procedure *(e.g. perforation, haemorrhage)*	E4. Problem knyttet til operativt inngrep *(f.eks. perforasjon, blødning)*	E4. Problems related to surgical intervention *(e.g. perforation, haemorrhage)*
	E5.Problem related to invasive procedure *(e.g. perforation, haemorrhage)*	E5. Problem knyttet til invasiv prosedyre *(f.eks. perforasjon, blødning)*	E5. Problems related to invasive procedure *(e.g. perforation, haemorrhage)*
	E6.Other procedure related problem	E6. Annet prosedyrerelatert problem	E6. Other procedure-related problems
F	Problem in clinical monitoring (including failure to plan, to undertake, or to recognise and respond to changes)	Problem innen klinisk overvåkning (inkludert manglende planlegging, utføring, eller erkjennelse av og respons på forandringer)	Problems in clinical monitoring (including insufficient planning, execution or perception of and response to changes)
	F1.Problem with monitoring TPR/BP/Sats/EWS	F1. Problem med monitorering av temperatur, puls, respirasjon, blodtrykk, O_2-metning	F1. Problems with monitoring of temperature, pulse, respiration, blood pressure, O_2 saturation
	F2.Problem with monitoring fluid intake/output	F2. Problem med monitorering av væskebalanse	F2. Problems with monitoring fluid balance
	F3.Problem with monitoring nutritional intake	F3. Problem med monitorering av næringsinntak	F3. Problems with monitoring nutritional intake
	F4.Problem with monitoring technology (e.g. telemetry)	F4. Problem med monitorering av teknologi (f.eks. telemetri)	F4. Problems with monitoring technology (e.g. telemetry)
	F5.Problem with monitoring neurological observations	F5. Problem med monitorering av nevrologiske observasjoner	F5. Problems with monitoring neurological observations
	F6.Problem with monitoring skin/wound condition	F6. Problem med monitorering av status i hud/operasjonssår	F6. Problems with monitoring of status of skin/operation wounds
	F7.Problem with monitoring via blood tests *(use only for routine monitoring)*	F7. Problem med monitorering vha. blodprøver *(bruk kun for rutinemessig monitorering)*	F7. Problems with monitoring by means of blood samples *(use for routine monitoring only)*

18

	F8.Other clinical monitoring problem	F8. Annet klinisk monitoreringsproblem	F8. Other clinical monitoring problems
G	Problem in resuscitation following a cardiac or respiratory arrest (including CPR)	Problem med gjenopplivning etter en hjerte- eller respirasjonsstans (inkludert HLR)	Problems with resuscitation after a cardiac or respiratory arrest (including CPR)
	G1.Delay in beginning resuscitation	G1. Forsinkelse med å sette i gang gjenopplivning	G1. Delay in initiating resuscitation
	G2.Problem in airway management	G2. Problem med sikring av frie luftveier	G2. Problems with safeguarding of unobstructed respiratory tract
	G3.Problem related to cardiac massage	G3. Problem knyttet til hjertemassasje	G3. Problems related to cardiac massage
	G4.Problem related to resuscitation medication/fluids	G4. Problem knyttet til gjenopplivningsmedisinering /-væskebehandling	G4. Problems related to resuscitation medication/fluid therapy
	G5.Problem related to resuscitation equipment	G5. Problem knyttet til gjenopplivningsutstyr	G5. Problems related to resuscitation equipment
	G6.Other resuscitation related problem	G6. Annet gjenopplivningsrelatert problem	G6. Other resuscitation-related problems
I	Any other problem not fitting categories above	Eventuelle andre problemer som ikke passer inn i kategoriene ovenfor	Any other problems that do not fit into any of the categories above
	I.Any other problem not fitting categories above *[Note free text option not required for any 'other' categories since the problem has already been described in free text]*	I. Eventuelle andre problemer som ikke passer inn i kategoriene ovenfor *[Merk: Fritekstvalgmulighet ikke påkrevet for noen 'annet'-kategorier, ettersom problemet allerede er blitt beskrevet i fritekst]*	I. Any other problems that do not fit into any of the categories above *[Note: Free text option not required for any of the 'other' categories because the problem has already been described in free text]*
	Apparent contributory factors to this problem in healthcare	**Tilsynelatende medvirkende faktorer til dette helsehjelpsproblemet**	**Apparent contributing factors to this health-care problem**
	Not possible to identify any clear contributory factors for this problem in healthcare from case note review	Ikke mulig å identifisere noen klare medvirkende faktorer til dette helsehjelpsproblemet fra journalgjennomgang og -vurdering	Not possible to identify any clear contributing factors to this health-care problem from a review and assessment of the case notes
	Patient factors *(e.g. did not disclose relevant history, ignored advice not to combine medication & alcohol)* Note do NOT use patient factors for comorbidities or confusion	Pasientfaktorer *(f.eks. fremla ikke relevant sykehistorie, ignorerte råd om ikke å blande sammen medikamenter og alkohol)* Merk: IKKE bruk pasientfaktorer for komorbiditeter og forvirring	Patient factors *(e.g. did not submit relevant case history, ignored advice about not taking medication together with alcohol)* Note: DO NOT use patient factors for co-morbidities and confusion
	Individual staff factors *(e.g. lack of insight into own competency*	Individuelle medarbeiderfaktorer *(f.eks.*	Individual employee factors *(e.g. insufficient*

	levels, failure to seek help, reckless, exhausted, ill)	*manglende innsikt i eget kompetansenivå, manglende evne til å søke hjelp, ubetenksom, utslitt, syk)*	*insight into own level of competence, insufficient ability to seek assistance, inconsiderate, worn out, sick)*
	Training, education and supervision factors *(e.g. lack of skills, knowledge or experience, lack of awareness of risks, inadequate supervision, no timely senior review)*	Opplærings-, utdannelses- og veiledningsfaktorer *(f.eks. manglende ferdigheter, kunnskap eller erfaring, manglende oppmerksomhet på risikoer, utilstrekkelig supervisjon, ingen gjennomgang og vurdering fra overordnet til rett tid)*	Training, education and guidance factors *(e.g. insufficient skills, knowledge or experience, insufficient attention to risks, inadequate supervision, no review and assessment by supervisor at the right time)*
	Task design, guideline and protocol factors *(e.g. misleading equipment design, outdated local protocol, missing or contradictory guidance)*	Oppgavedesign-, retningslinje- og protokollfaktorer *(f.eks. villedende utstyrsdesign, utdatert lokal protokoll, manglende eller motstridende veiledning)*	Task design, guideline and protocol factors *(e.g. misleading equipment design, outdated local protocol, insufficient or contradictory guidance)*
	Teamwork, leadership and communication factors *(e.g. inadequate handover, inadequate inter-professional challenge, weak leadership, missing paper or electronic records)*	Teamarbeids-, ledelses- og kommunikasjonsfaktorer *(f.eks. mangelfull rapport, mangelfull tverrfaglig utfordring, svak ledelse, manglende papirbaserte eller elektroniske journaler)*	Teamwork, management and communication factors *(e.g. insufficient delivery, insufficient inter-disciplinary challenge, weak management, insufficient paper-based or electronic case notes)*
	Local work environment factors *(e.g. noise, distractions, inadequate staffing, resources, missing equipment)*	Lokale arbeidsmiljøfaktorer *(f.eks. støy, forstyrrelser, underbemanning, ressursmangel, manglende utstyr)*	Local working environment factors *(e.g. noise, disturbances, understaffing, insufficient resources, insufficient equipment)*
	Organisation-wide factors *(e.g. overall bed capacity, design of IT systems, recruitment freezes)*	Faktorer som berører hele organisasjonen *(f.eks. samlet sengekapasitet, utforming av IT-systemer, ansettelsesstopp)*	Factors that affect the whole organisation *(e.g. total bed capacity, design of IT systems, hiring freeze)*
	Other type of contributory factor ……	Annen type medvirkende faktor ……	Other types of contributing factors .....

Vedlegg 4 Antagelser og paradokser fra Zhao, Liu og Deng (2013) (forenklet).

Antagelser	Nummer	Cohens Kappa	Gwets AC1	% Enighet
Definisjon av tilfeldig enighet	1,2	maksimal	maksimal	Ingen
Oppriktig skåring	3	Begrenset	Begrenset	Total
Spesiefisert tilfeldighet	4	Ja	Ja	Nei
Runder med "klinkekuletrekk"	23	En	To	Ingen
Tilbakelegging etter trekning	7,18	Ja	Ja	Ikke relevant
Klinkekulemønster = oppriktighet	8,22,24	Ulike	Ulike eller 2 like	Ikke relevant
Kategorier =farger	5	Ja	Ja	Nei
Likt nummer kuler per farge	6	Nei	Ja	Nei
Kategorier reduserer IRR	9	Nei	Nei	Nei
Enighet observeres eller anslås	10	Observert	Observert	Observert
opphøyet indeks	11	Nei	Nei	Nei
Marginalfordeling	12,17	Individuell	Sammenslått	Nei
Treenighetsantagelse	13	Ja	Ja	Nei
Begrenset oppgave	14	Ja	Ja	Nei
Forutbestemt fordeling	15	Ja	Ja	Nei
Fordeling påvirker IRR	16	Ja	Ja	Nei
Treenighet størrelse	19	Nei	Nei	Nei
Forutbestemt størrelse	20	Nei	Nei	Nei
Større utvalg øker påvirker	21	Nei	Nei	Nei

Paradokser	**Nr.**	***Cohens kappa***	***Gwets AC***	***% Enighet***
Gjetting er reliabelt	1			X
Ingenting annet enn tilfeldighet	2	X	X	
Epler kan sammenliknes med appelsiner	3	X	X	
Mennesker er en undergruppe av menn	4	X	X	
Pandaer er en undergruppe av menn	5	X	X	
Økende antall kategorier øker reliabilitet	6*		X	
Abnormaliteter				
Høy enighet og lav reliabilitet	10	X		
Udefinert reliablitet	11	X		
Lik observert enighet, stort fall i r	12	X		
Ingen uenighet, ingen økning i reliabilitet	13	X		
Lite økning i enighet, stor økning i reliabilitet	14	X		
Økt observ. Enighet, stort fall i r	15	X		
Oppriktig koding tilsvarer myntkast	16	X		
Bedret koding straffer seg	17	X		
Enighet straffer seg	18	X		
Terskelen flytter seg	19	X	X	
Sirkulær logikk	20	X	X	
Samme kvalitet og observerte enighet, høyere r	21		X	
lavere kvalitet og bservert enighet, høyere r	22		X	

*Paradoksene 7,8 og 9 angår ikke Cohens Kappa eller Gwets AC

Variabel 21 Generell Helsehjelpskvalitet

Gransker A fant at den helsehjelpskvaliteten var utilstrekkelig i 2.5% av journalene (n=5), og granskers B fant det samme i 10.5% av journalene (21). Det var enighet om at helsehjelpskvaliteten var utilstrekkelig for 0.5% av journalene (n=1). Det var enighet om at helsehjelpskvaliteten var tilstrekkelig for 87.9% av journalene (n=175). Granskerne var uenige i 11.5% av journalene (n=22)

Observert enighet var 88.9%, Kappa var.118, som tolkes kvalitativt som *slight,* med *n*edre konfidensintervall innen tolkningen poor (-.066), tilsvarende dårligere reliabilitet enn tilfeldig enighet. Gwets AC1 var .874 som tolkes som *nesten perfekt,* med nedre konfidensintervall innen samme kateogir (.819).

Tabell : Generell helsehjelpskvalitet

		Gransker B		
		Tilstrekkelig	*Utilstrekkelig*	Total
Gransker A	*Tilstrekkelig*	175	19	194
	Utilstrekkelig	3	2	5
	Total	178	21	199

Nr	Variabelnavn	% Enighet	Kappa	KI		Sig.	AC1	KI		Sig.	Kappa	Gwets AC
#	Kvalitet av helsehjelpen	88.9%	.118	-.066	.302	<.05	.874	.819	.929	<.001	Slight	Almost perfect

Variabel 22 Inngripende og uhensiktsmessige prosedyrer

tiltak Gransker A fant at det hadde forekommet inngripende eller uhensiktsmessige tiltak hos 3% av pasientene (n=6), og granskers B identifiserte slike tiltak hos 10.5% (n=21). Det var enighet mellom granskerne om at det hadde forekommet inngripende eller uhensiktsmessige tiltak hos 0,5% (n=1), og enighet om at det ikke hadde funnet sted slike tiltak hos 81%. (n=162) Granskerne var uenige i 12.5% av journalene (n=25)

For ett av subjektene fant gransker B at det forelå en belastende prosedyre, mens gransker A ikke trakk slutning. Ettersom usikkerhetskategorien slettes og blir en manglende verdi, slettes disse to skåringene parvis. De øvrige parvise slettingene var negative funn.

Observert enighet var 86,6%, Kappa var .026, som tolkes kvalitativt som *Slight,* med *n*edre konfidensintervall innen tolkningen poor (-.111) Kappa var ikke signifikant høyere enn 0, (*P*- verdi .664), noe som innebærer at beregningen ikke er vesentlig forskjellig fra reliabilitet ved tilfeldig skåring. Gwets AC1 var .846 som tolkes som *Almost perfect,* med nedre konfidensintervall innen tolkningen *Substantial* (.783).

Tabell: Inngripende og uhensiktsmessige prosedyrer

Manglende verdier n= 12, 6%

		Gransker B		
		Ja	*Nei*	Total
Gransker A	*Ja*	1	5	6
	Nei	20	162	182
	Total	21	167	188

Nr.	Variabelnavn	% Enighe	Cohens Kappa	KI		p	Gwets AC1	KI		Sig.
22**	Inngripende og uhensiktsmessige prosedyrer Ja/nei	86.7%	.026	- .111	.163	.664	.846	.783	.910	<.001

Variabel 29 Dokumentasjonskvalitet, Opprinnelig ordinal variabel

Cohens Kappa (vektet) for variabelen var ,102 (*p*-verdi <.05) som gir kvalitativ tolkning *Slight.*

Wilcoxon signed Rank: Lege B vurderte dokumentasjonsgrunnlaget som bedre enn A i 63 journaler, og dårligere i 23 journaler. Legene var enige i 115 av journalene.
Forskjellene i vurderingene av dokumentasjonsgrunnlaget var signifikant (p<.005) og effektstørrelsen var (r = z/(√200)) =(-) 0,21 : Liten effekt (<(-)0,3)

Shapiro wilk:
Lege A: W = .561
Lege B: W= .475
Begge: p <.05

Dokumentasjonskvalitet, dikotmisert

Gransker A fant at dokumentasjonskvaliteten var tilfredsstillende for alle journaler (n=188)
Gransker B fant at dokumentasjonskvaliteten var tilfredsstillende i 185 journaler (98.4%). Det var enighet mellom granskerne for alle journalene gransker B vurderte som tilfredsstillende (98.4%, n=185)

Observert enighet for dokumentasjonskvalitet var 98.4%, Kappa kunne ikke beregnes, ettersom gransker A skåret tilstrekkelig dokumentasjonskvalitet for alle journaler. Gwets AC1 var .984, som tilsvarer prosent enighet og tolkes som *Almost perfect,* med nedre konfidensintervall (.965) innen samme tolkning

Manglende verdier n=2, 1%

		Gransker B		
		Tilstrekkelig	*Utilstrekkelig*	Total
Gransker A	*Tilstrekkelig*	185	3	188
	Total	185	3	188

			Cohens			Gwets			
Nr.	Variabelnavn	% Enighet	Kappa	KI	p	AC1	KI		Sig.
29	Dokumentasjonskvalitet Tilstrekkelig/ utilstrekkelig	98.4%	**			.984	.965	1	<.000

Tidsbruk

Shapiro-Wilk test gir p<.001 for begge gjennomganger og begge grupper, visuelt (boxplot)er begge grupper høyreforskjøvet og har flere høye outliere.
Lager normalfordelte (Log10 variabler og kontrollerer at de er normalfordelte (p >.05) 7
Utfører t- test som viser at det er en ikke- signifikant forskjell mellom legegranskernes mean for tidsbruk utvalget som helhet og for gruppen menn (p>.05). For gruppen kvinner er en signifikant forskjell i mean(p<.05) og forskjellen er liten til medium stor (d= .286, df =81)

Point biserial korrelasjon:

Lege A

signifikante korrelasjoner (p<.01) mellom tidsbruk og helsehjelpsproblem for begge kjønn(r=.492), for menn (r=469) og for kvinner (r=. 516).
signifikante korrelasjoner (p<.01) mellom tidsbruk og forebyggbare dødsfall for begge kjønn(r=.539), for menn (r=384) og for kvinner (r=. 459).

Lege B

Også signifikante korrelsjoner (p= <.01) mellom tidsbruk og helsehjelpsproblemer for utvalget (r=.445), for menn(r=444), og for Kvinner (.439).
Signifikante korrelasjoner (p<.01) mellom tidsbruk og forebyggbare dødsfall for begge kjønn (r=.453), Menn (r=523),og kvinner (r=.349)

Visuelt er korrelasjonene positive, større tidsbruk øker identifikasjon av helsehjelpsproblemer

Logistisk regresjon for dødsfall

Lege A:

42,7% % av vurderingene av forebyggbare dødsfall kan forklares av tidsbruk (p<.001), og for hvert minutt økning i tidsbruk øker risikoen for å få en vurdering av forebyggbare dødsfall med 6 % (p<.001)

2,7 % av unngåelige dødsfall hos menn og 2,2 % hos kvinner kan forklares av variasjon i alder. For hvert år aldring faller odds for å rammes av unngåelig dødsfall for menn med 2% og for kvinner øker risiko med 1%. Resultatene er ikke signifikante (P>0,05)

Lege B

30,3 % av vurderingene av forebyggbare dødsfall kan forklares av tidsbruk (p<.001), og for hvert minutt økning i tidsbruk øker risikoen for å få en vurdering av forebyggbare dødsfall med 4 % (p<.001)

7 % av unngåelige dødsfall hos menn og 4% hos kvinner kan forklares av variasjon i alder. For hvert år aldring øker unngåelig dødsfall for menn med 3% og for kvinner øker risiko med 2% Resultatene er ikke signifikante (P>0,05),men nesten signifikante for menn (p=0,051)