Segmentering av cerebrospinalvæske på MR-bilder av barn mellom 0 og 2 år

Utvikling og evaluering av metode

Ellen Regine Olsrud

HELSEF4502: Masteroppgave i interdisiplinær helseforskning

Universitetet i Oslo
Det medisinske fakultet
Institutt for helse og samfunn
Avdeling for tverrfaglig helsevitenskap

Mai 2021

2021

Segmentering av cerebrospinalvæske på MR-bilder av barn mellom 0 og 2 år

Utvikling og evaluering av metode

http://www.duo.uio.no

Trykk: Reprosentralen, Universitetet i Oslo

Forord

Masterstudien ble gjennomført ved Avdeling for tverrfaglig helsevitenskap, Universitetet i Oslo, og Oslo universitetssykehus, fra august 2018 til mai 2021.

Våren 2018 fikk jeg muligheten til å bli med i et prosjekt som innebar segmentering av CSF i MR-bilder ved forskning og utviklingsavdeling (FOU), Ullevål, OUS. Denne unike muligheten har gitt et bredt spekter av ny kunnskap. En stor takk til min hovedveileder, overlege og professor II Heidi B. Eggesbø, for at jeg fikk denne tilliten og for all veiledning gjennom masteroppgaven. Jeg har satt stor pris på ditt ustoppelige engasjement for prosjektet og masteroppgaven.

En stor takk rettes og til min nåværende leder, PhD Frode A. Tuvnes, både for kunnskapsrike og konstruktive innspill, men og for din herlige humor som gjør hverdagen en del lysere. En like stor takk til tidligere leder, Wenche Synnøve Andreassen, for at jeg fikk tid til å holde på med dette, i en travel hverdag på seksjon for MR.

Jeg er svært takknemlig for dyktig veiledning fra Hilde S. Robinson, førsteamanuensis ved UiO, din statistikk-ekspertise og blikk fra «utsiden» har vært gull verdt.

Nevroradiolog Maninder Singh Chawla, tusen takk for all kunnskapen du velvillig deler og alle timene du tålmodig guidet oss gjennom MR-bilder snitt for snitt. Tusen takk til nevroradiolog Paul Debrah Karikari, for hjelp i oppstartsfasen med grundig veiledning i anatomi.

En like stor takk til fysiker Wibeke Nordhøy, for gjennomlesing, og til fysiker Robin A. Bugge for innspill, og til vår hjelpsomme AI-utvikler og radiolog Tomas Sakinis, uten dine ferdigheter, hadde vi ikke kommet der vi er i dag.

Til min samarbeidspartner, MR-radiograf og venn, Bianca Lund-Melcher, det er en fornøyelse å jobbe på lag med deg, takk for ditt bidrag til datainnsamling. Og til Lisa Kjønigsen, tusen takk for korrekturlesing og for å ha bidratt til lavere puls når stormen har rast.

Mange flere fortjener takk, men viktigst av alle: min herlige familie som med god grunn har telt ned dagene til innlevering.

Sammendrag

Bakgrunn

En debatt om hvorvidt tilstanden godartet vannhode (BEH) kunne gi samme symptombilde som ved «filleristing» av spedbarn, synliggjorde et behov for mer viten om volumet av cerebrospinalvæske (CSF) hos barn gjennom de to første leveårene. Implementeringen av kunstig intelligens (KI) i radiologi har gitt muligheter for å utvikle algoritmer til å utføre automatisk segmentering. Vi startet derfor en studie for å utvikle en KI-basert metode for å måle CSF volum ved segmentering av MR-bilder.

Formål

Hensikten med masterstudien var å utvikle en KI-metode for segmentering av ventrikkel CSF og subaraknoidal CSF på MR-bilder av barn fra 0 til 2 år, og evaluere metoden med hensyn til validitet og reliabilitet.

Metode

Aksiale T2 vektede MR-bilder ble først segmentert manuelt for ventrikkel CSF og subaraknoidal CSF. De segmenterte bildene ble deretter brukt til å trene opp en dyplæringsalgoritme for automatisk segmentering. Tilsammen fem algoritmeversjoner med teoretisk økt presisjon ble utviklet. Samsvar for volum og piksel-overlapp ble evaluert mellom automatiske segmenteringer og manuelt korrigerte gullstandard for validering av algoritmen. Inter-rater reliabilitet ble utført ved to ulike grupper, en med normale og en med økte CSF volum. Intra-rater reliabilitet utført i en gruppe med normale CSF volum.

Resultat

Piksel-overlapp (Dice koeffisient) og volum (ml) viste høyt samsvar for de to siste KI-algoritme versjonene versus gullstandard, både for ventrikkel og subaraknoidal CSF (Dice-koeffisient $\geq$ 0,97, ICC = 1,000). Inter- og intra-rater reliabiliteten var høy (Dice-koeffisient $\geq$ 0,96, ICC = 0,997), men noe lavere for barn med økt CSF volum.

Konklusjon

I denne studien har vi utviklet og evaluert en metode basert på kunstig intelligens, for segmentering av ventrikkel CSF og subaraknoidal CSF. Den validerte KI metoden kan bli brukt for segmentering av et større materiale, med hensikt om å danne referansemateriale, for CSF volum gjennom de to første leveårene.

Abstract

Background

A debate about whether the condition benign external hydrocephalus (BEH) could present the same symptoms as in abusive head trauma in children, highlighted a need for more knowledge about the cerebrospinal fluid (CSF) volume in children. The implementation of artificial intelligence (AI) in radiology has made it possible to develop algorithms for automatic segmentation. Therefore, we started a study in order to develop an AI-based method for magnetic resonance imaging (MRI) segmentation of CSF in children.

Objectives

The aim of this study was to develop a method for segmentation of CSF volume in children 0-2 years, and then evaluate the method in terms of validity and reliability.

Methods

Axial T2 weighed MRI were first manually segmented for ventricular CSF and subarachnoid CSF. The MRI-segmentations were used to train a deep learning algorithm. Five versions with theoretically increased precision were developed. Volume and pixel overlap were evaluated between automatic segmentation and manually corrected god standards for validation of the algorithm. Inter-rater reliability was examined in two different groups, one with normal and one with increased CSF volume. Intra-rater reliability was performed in one group with normal CSF volume.

Results

Both volume and pixel overlap (Dice coefficient) showed high agreement between the two last versions of the AI algorithm and the gold standard for both ventricular and subarachnoid CSF (Dice-coefficient ≥ 0,97, ICC = 1,000). Further, inter- and intra-rater reliability were high (Dice-coefficient ≥ 0,96, ICC = 0,997), but slightly lower in children with increased CSF volume.

Conclusion

In this study, we have developed and evaluated an AI-based method for segmentation of ventricular CSF and subarachnoid CSF in children aged 0-2 years. The validated method can be used for segmentation of a larger material in order to make a reference values for CSF volume during the first two living years.

Forkortelser

2D	to-dimensjonal
3D	tre-dimensjonal
CSF	cerebrospinal fluid / cerebrospinalvæske
CT	computer tomografi
DICOM	Digital Imaging and Communications in Medicine
DSC	Dice similarity coefficient
GE	gradient ekko
ICC	intraklasse korrelasjons koeffisient
KI	kunstig intelligens
KNN	konvolusjonelle nevrale nettverk
KRN	Klinikk for radiologi og nukleærmedisin
LoA	limits of agreement
OUS	Oslo universitetssykehus
MR	magnetisk resonans
PVE	partiell volumeffekt
RF	radiobølger
sCSF	subaraknoidal CSF
SE	spinn ekko
SD	standardavvik
SEM	standard målefeil / standard error of measurement
T1	T1 vektet sekvens
T2	T2 vektet sekvens
vCSF	ventrikkel CSF
QQ plot	quantile-quantile plot

Figurer

Tabeller

Innholdsfortegnelse

1. Bakgrunn

I 2017 publiserte svenske 'Statens Beredning för medicinsk och social utvärdering' (SBU) en utredning om triaden av symptomer som ses ved såkalt "filleristing" av spedbarn (1). Triaden innebefatter blødninger i netthinnen, blødning under hjernens harde hinne (subduralblødning) og encefalopati (diverse hjernelidelser). Rapportens konklusjon var at det var begrenset vitenskapelig bevis for at triaden var forbundet med «filleristing» og at man ikke kunne identifisere påført skade på bakgrunn av triaden. Rapporten utløste debatt og ble sterkt kritisert og beskyldt for å redusere rettssikkerheten til barn utsatt for vold (2, 3).

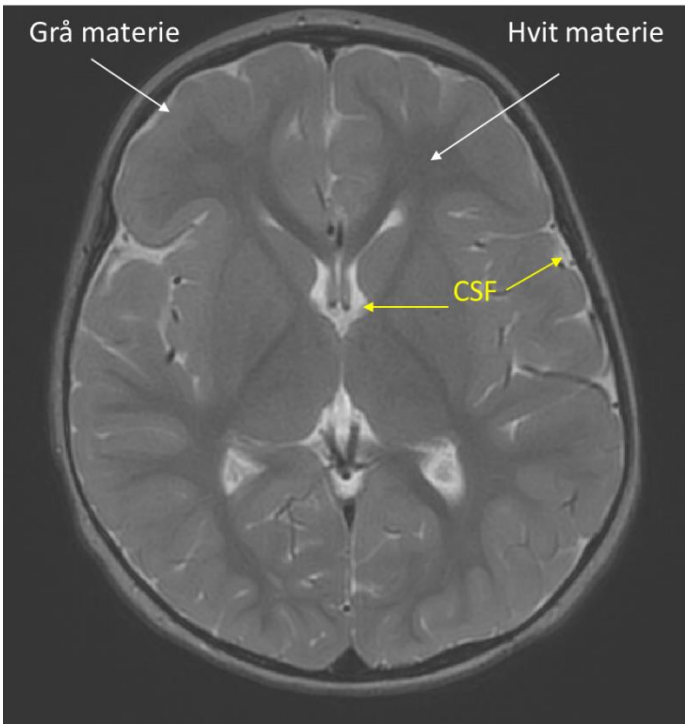
En norsk pensjonert professor i nevrokirurgi som delte oppfatning med den svenske rapporten, mente symptomene i triaden også kunne forklares med tilstanden godartet ytre vannhode, benign ekstern hydrocephalus (BEH) (4). BEH har vært en differensialdiagnose i mer enn 20 år (3). BEH medfører økt hodeomkrets, med økt volum av cerebrospinalvæske (CSF) i subaraknoidalrommet, mens ventriklene er tilnærmet normale. Tilstanden blir vanligvis ikke behandlet, da den ofte går over av seg selv ved 2 års alder (5, 6). Det er holdepunkter for at tilstanden kan føre til utsiving av blodprodukter i subaraknoidalrommet (7). BEH er forskjellig fra hydrocephalus, som oftest er behandlingskrevende. Hydrocephalus skyldes enten hinder i CSF sirkulasjonen eller redusert absorpsjon, og fører til økt intrakranielt trykk og forstørrede ventrikler (8).

Denne debatten synliggjorde et behov for mer kunnskap om CSF volum i barns første leveår. Per i dag finnes det ingen verdier for CSF volum hos barn fra 0 til 2 år. Med dette som bakgrunn startet vi en studie med mål om å lage et referansemateriale og metode for å måle CSF volum hos barn fra 0-2 år.

Med bakgrunn som erfarne MR radiografer deltok jeg og en kollega i studien for å utføre målingene av CSF volum, ved segmentering av MR-bilder. Vi fikk grundig veiledning fra spesialist i pediatrisk nevroradiologi. Før et referansemateriale kunne dannes, måtte metoden utvikles og evalueres.

1.1 Introduksjon

I løpet av de siste tiårene har utviklingen innen medisinsk bildeteknologi, som magnetisk resonans (MR), ført til ny kunnskap om hjernens anatomi og fysiologi (9). MR-avbildning gir en unik fremstilling av hjernen, der man kan klassifisere tre hovedgrupper av vevstyper, hvit materie som hovedsakelig består av myeliniserte nerveforbindelser, grå materie som hovedsakelig består av cellekropper og støtteceller samt CSF, som vist i Figur 1. CSF er en væske som omslutter hele sentralnervesystemet, og kan deles i ventrikkelsystemet og subaraknoidalrommet (10).



Figur 1 MR bilde av hjernen hos et barn på 24 måneder: tverrsnittplan (aksialt), hvit materie fremstilles her i mørkere gråtoner enn grå materie, CSF har høyere signal og er hvit. Bildeopptaket, sekvensen, er T2 vektet. Bildet tilhører studiens datasett.

Radiologisk bildediagnostikk med MR og computertomografi (CT), inneholder informasjon som kan gi volumdata av vevstyper, men har i liten grad vært en del av den kliniske vurderingen av bilder (11). Når radiologer studerer bildene og vurderer CSF volum, er dette i stor grad en kvalitativ og subjektiv vurdering. Radiologene kan som oftest bekrefte økt CSF volum, uten at de har eksakt volummål. Informasjonen er der, men for å kvantifisere volumet, må bildene segmenteres.

Segmentering innebærer å dele opp bildet i regioner etter egenskaper som gråtoner, for å studere anatomi, identifisere lesjoner eller måle volum (12). Segmentering kan være manuell, eller automatisk ved en algoritme. Segmenteringsalgoritmer er dataprogrammer som analyserer bildene etter en innlært oppskrift, og kan gi volum av hjernens strukturer og CSF. Slike algoritmer som er basert på kunstig intelligens (KI), har gitt unike muligheter for å hente ut kvantitative data, volum, fra MR og CT undersøkelser.

Bruk av segmenteringsalgoritmer basert på KI ble en viktig faktor for studien der vi skulle måle CSF volum. Denne masterstudien omhandler metodeutvikling for segmentering av CSF i MR-bilder, og evaluering av validitet og reliabilitet til metoden.

1.2 Begrepsavklaringer

Datasett: hvert datasett består av 20-40 MR bilder av en hjerne fra foramen magnum, bunnen av kraniet, til toppen av hodet. Et datasett genererer en variabel for vCSF og en variabel med sCSF.

Snitt: tilsvarer ett bilde

Sekvens: tilsvarer et MR opptak, en serie, e.g. T2 vektet sekvens

T2 vekting: MR sekvens med høyt signal fra væske, som blir hvit og lys, lavere signaler fra grå og hvit substans, som blir ulike gråtoner.

Segmentering: prosess der en vevstype eller struktur skilles ut og markeres. Denne markeringen kan kvantifiseres som et volum.

Gullstandard: tilsvarer en felles forståelse og kriterier for optimal CSF segmentering. Gullstandard brukes i denne studien som begrep for datasett som er segmentert med KI-algoritme, med etterfølgende manuell korrigering/segmentering.

KI-algoritmen: brukes om segmenteringsalgoritmen som er basert på kunstig intelligens (KI), utviklet for studien og trent opp med studiens datasett.

KI-algoritme versjon 1-5: Opptrening av algoritmen genererer en ny versjon for hver runde av trening. Opptreningen gjøres med segmenterte datasett på nivå med gullstandard. Hver ny versjon antas å ha økende presisjon i utførelsen av segmenteringen.

Manuell korrigering: segmentering som utføres i dataprogrammet SliceOmatic. Innebærer å rette opp i feilsegmenteringer utført av KI-algoritmen.

2. Teori

I dette kapittelet vil cerebrospinalvæske (CSF), modaliteten (MR), segmentering og kunstig intelligens (KI) bli redegjort for.

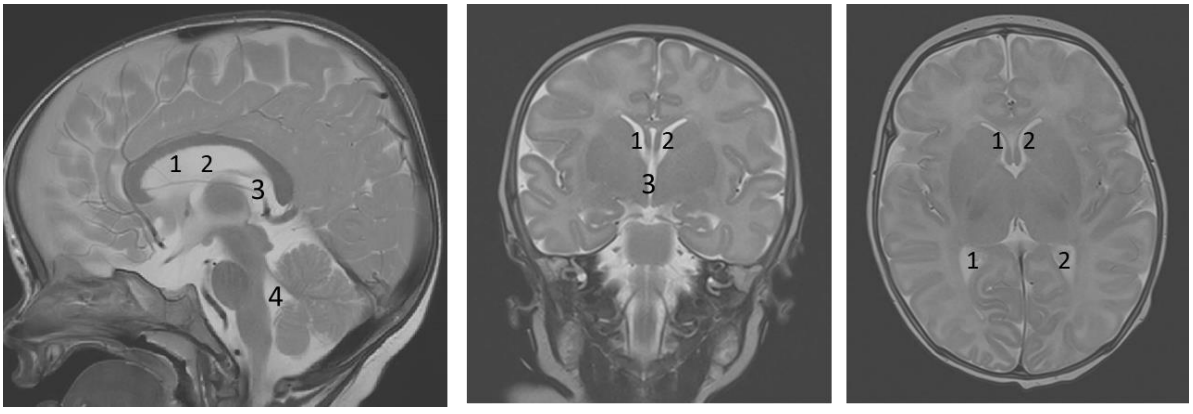
2.1 CSF og hjernen

I løpet av de to første leveårene skjer det en enorm utvikling av hjernen. Gjennom første leveår dobles hjernestørrelsen, og ved 2 års alderen er den 80 % av en fullvokst hjerne (13). CSF volumet øker i takt med hjernen (10).

CSF er væsken som ligger i hjernens hulrom, ventriklene, samt rundt hjernen og ryggmargen, som kalles subaraknoidalrommet (14). Ventrikkelsystemet består av fire ventrikler: lateralventriklene, tredje og fjerde ventrikkel, vist i Figur 2. Heretter blir ventrikkel CSF referert til som vCSF, og subaraknoidal CSF referert til som sCSF.

Gjennomsnittlig CSF volum for voksne er 150 ml, av dette er ca. 25 ml er i ventriklene (10). Væsken fornyes ca. fire ganger per døgn, og en stor del av produksjonen skjer i choroid plexus i sideventriklene og i tela choroidea i tredje og fjerde ventrikkel (10). Den arterielle pulsasjonen påvirker CSF væskens bevegelser (15).

CSF beskytter hjernen mot slag/støt, bidrar til utveksling av næringsstoffer mellom blodårer og hjernevev, og bidrar som transportmedium for utskillelse av avfallsstoffer (16). I 2017 satte Ringstad et al. utvaskingen av avfallsstoffer i sammenheng med søvn og demonstrerte økt «hjernevask» i forbindelse med søvnstadier ved hjelp av MR skanninger (17).



Figur 2 MR bilder av CSF fremstilt i tre plan: sagittalt til venstre, koronalt i midten og aksialt til høyre: CSF fremstilles hvit/høyt signal. vCSF befinner seg i ventriklene markert i tall: 1. – 4. ventrikkel, i hjernens hulrom. sCSF befinner seg i subaraknoidalrommet som ligger rundt hjernen og ryggraden. Bildene tilhører studiens datasett.

2.2 MR-avbildning

MR er den foretrukne radiologiske modaliteten for bilde-diagnostikk av hjernen. MR-maskinen har et kraftig statisk magnetfelt, hvor styrken oppgis i Tesla. 1 Tesla er det samme som 10 000 Gauss, som er 20 000 ganger sterkere enn jordmagnetfeltet på 0,5 Gauss (18). For nevroradiologiske (hjerne og ryggmarg) undersøkelser er 1.5 og 3 Tesla mest brukt.

MR-maskinen består av, i tillegg til det kraftige magnetfeltet, en radiobølge (RF) sendespole som sender ut et signal som lokaliseres av tre gradientspoler/elektromagneter i x-, y- og z-plan i kroppen, og en mottakerspole som leser ut signalet (19).

Det er protonene i hydrogenatomene som er signalgivende på MR. Det statiske magnetfeltet fører til at flertallet av protonene retter seg inn langs med det parallelle magnetfeltet. For å få et signal, sendes det inn en radiobølge, i form av RF pulser, som kommer i resonans med de eksiterte protonene og får de ut av likevekt. Når RF pulsen slås av, vender de tilbake til likevekt (19). Dette kalles longitudinell relaksasjon (T1 relaksasjon). Samtidig foregår det også tap av transversal vevsmagnetisering (T2 relaksasjon). Relaksasjonsprosessene er forskjellige i ulike typer vev, og får ulike gråtoner i MR-bildene basert på blant annet av når man leser ut signalet. Vann (som CSF) har lang relaksasjonstid, og avgir dermed høyt signal i T2 vektede bilder, og blir lyst eller hvitt i bildene (19).

En MR-undersøkelse består av et varierende antall bildeopptak også kalt sekvenser, som velges ut fra en klinisk problemstilling. Hver sekvens har en type vekting i forhold til timing av T1- og T2-tidene ved signalutlesningen, som gir ulike kontraster i bildene. For nevroradiologiske undersøkelser er de mest brukte sekvensene: T1, T2 og diffusjonsvekting (19).

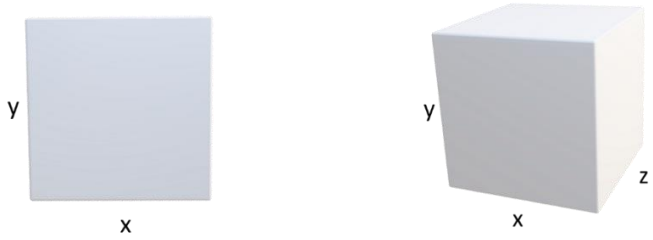
Det skilles i hovedsak mellom spinn ekko (SE) og gradient ekko (GE) sekvenser, hvor forskjellen er om det er RF-pulser eller feltgradienter som refokuserer protonene før signalutlesningen. For å få en ren T2-vekting, må man bruke varianter av SE-sekvenser (19).

MR-sekvenser er enten to-dimensjonale (2D), hvor ett og ett snitt med en viss snittykkelse tas opp hver for seg, eller tre-dimensjonale (3D), som er et volumopptak som kan rekonstrueres i ønskede plan etterpå, vanligvis som aksial, sagittal og koronal (18).

I motsetning til andre radiologiske undersøkelser som bruker ioniserende stråling som konvensjonell røntgen og computer tomografi (CT), er det ingen kjent risiko ved bruk av MR som er ikke-ioniserende og hvor magnetfeltpåvirkningen er reverserbar (18).

Piksler og voksler i et MR-bilde

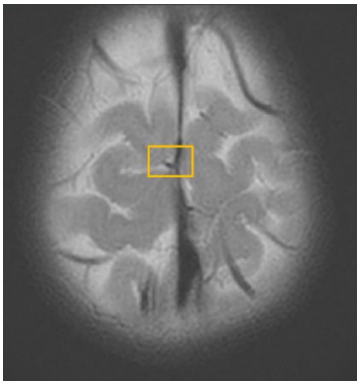
Et bilde fra en MR-sekvens består av et absolutt antall bildeelementer, hvor to-dimensjonale bilder består av piksler, og tre-dimensjonale bilder viser voksler. Hver piksel i bildet representerer en signalverdi og kan direkte relateres til en voksel (18). Pikslene har en x- og y-dimensjon som til sammen utgjør et areal, og vokslene har tre dimensjoner, x, y og z, som utgjør et volum (9).



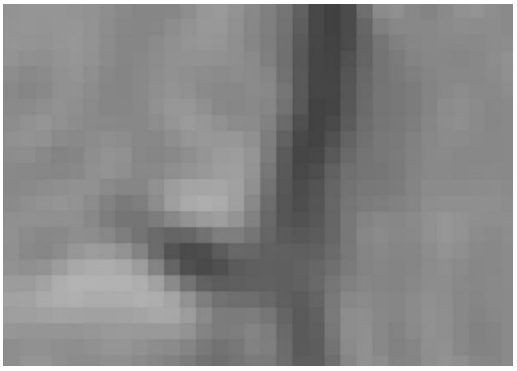
Piksel *Voksel*

Figur 3 Illustrasjon av en piksel og en voksel. En voksel har en ekstra dimensjon i forhold til en piksel.

Antall piksler i bildet, definerer bildematrisen, som kan ses på som et rutenett. Bildematrisen sier noe om hvor mange elementer det er i x- og y-retning. Oppløsningen i bildet, refererer til piksel- og voksel-størrelsen, eksempel vist i Figur 4. Høy oppløsning tilsvarer små piksler og voksler.



Figur 4a



Figur 4b *Forstørret (gult rektangel) fra Figur 4a*

Figur 4a og 4b Piksler i MR utsnitt: a) viser aksialt snitt i øvre del av hjernen, b) er forstørret utsnitt. Illustrasjon av eksempel på matrise og oppløsning i et MR-bilde. CSF fremstår hvitt, blodårer fremstår svart og grå materie er fremstilt i ulike gråtoner. Bildet tilhører studiens datasett.

Segmentering av MR-bilder

Segmentere betyr å skille ut og måle, og er en godt adaptert metode innen medisinsk forskning (9). I hovedsak er det to måter å segmentere, manuelt eller automatisk. Automatisk segmentering skjer via et dataprogram, en algoritme, manuell segmentering utføres ved å markere piksler i MR-bildet. Automatisk segmentering gir resultater nærmest umiddelbart i motsetning til manuell segmentering som er svært tidkrevende (20). Manuell segmentering utført av eksperter, regnes som gullstandard, forstått som fasit, men er ofte for tidkrevende til at det er egnet for større datasett (21).

Det finnes mange tilgjengelige automatiske segmenteringsprogrammer for MR bilder av hjernen, men ulempen er at de i hovedsak baseres på voksne hjerner og/eller krever spesifikke sekvenser med høy oppløsning (22). Segmentering av hjerner hos spedbarn er mer komplisert grunnet betydelige anatomiske utviklingen og ofte redusert kvalitet av MR-undersøkelsen sammenlignet med voksne (9, 23).

Artefakter

Artefakter betyr at bildet fremstiller noe som ikke er i objektet, enten i form av støy eller misvisende signaler (19). Relevant for segmentering er bevegelsesartefakter, som skyldes bevegelser fra pasienten, strømnings (flow)-artefakter som skyldes pulsasjoner, og partiell volumeffekt, som skyldes at flere typer vev inngår i samme piksel eller voksel (12, 13).

Barn under 6 måneder er ekstra utsatt for bevegelsesartefakter da de hovedsakelig ikke sederes ved MR undersøkelser, de immobiliseres ved puter/tøy og sukkervann. Flow-artefakter skyldes pulsasjoner fra arterier og CSF, og kan føre til signaltap i bildet. Piksler som tilhører CSF kan dermed gi lavt signal, det vil si svarte, på T2 sekvenser (19).

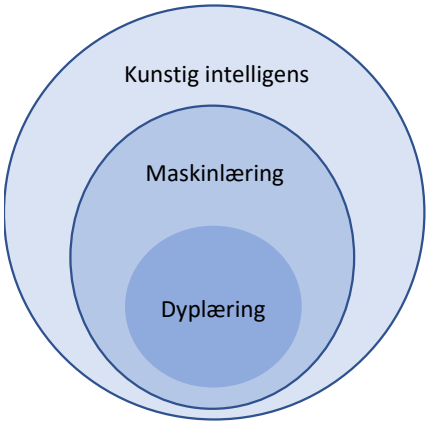
Barnehjerner utgjør et mindre volum og krever høyere oppløsning i forhold til en voksen hjerne (23). Høyere oppløsning øker opptakstiden eller fører til mer støy. Opptakstiden er begrenset ettersom lengre sekvenser øker sannsynligheten for bevegelse. En 2D sekvens varer som regel 3-5 minutter og en fullstendig undersøkelse krever som regel fire eller flere ulike sekvenser. Tidsbegrensning gjør at man ikke alltid får optimal oppløsning.

Små volum og suboptimal oppløsning gir økt forekomst av partiell volumeffekt (PVE), og gjør CSF segmentering ekstra utfordrende (11, 24). PVE oppstår når ulike typer vev, som hjerne og CSF, inngår i samme piksel/voksel, som vist i Figur 5. Resultatet er at pikslene og vokslene får en gjennomsnittsgråtone (19).

Figur 5 **Illustrasjon av partiell volumeffekt.** Utsnitt fra en aksial T2 sekvens: i oransje firkant ses mellomgrå piksler med diffuse avgrensinger, årsaken er voksler hvor både CSF og hjernevev er avbildet. Bildet tilhører studiens datasett.

2.3 Kunstig intelligens, maskinlæring og dyplæring

Kunstig intelligens (KI) er et generelt begrep om hvordan dataprogrammer etterligner intelligent adferd med minimal menneskelig intervensjon (25). KI i radiologi har hittil ført til bedre utnyttelse av bildeopptak og avansert postprosessering (26). Postprosessering betyr etterarbeid av bilder, og inngår i klinisk diagnostikk, terapi, og forskning. Dette gjør at mer informasjon kan hentes ut av bildene. KI er spesielt egnet for å påvise og klassifisere lesjoner, dataanalyser, bilderekonstruksjon og automatisk segmentering (26). Forholdet mellom kunstig intelligens, maskinlæring og dyplæring, kan fremstilles som i Figur 6 (27):



Figur 6 Kunstig intelligens og undergruppene maskinlæring og dyplæring.

Maskinlæring og dyplæring er felt innen kunstig intelligens, der algoritmer trenes opp for å finne mønstre i data. Begge bruker kunstige nevrale nettverk som er analytiske algoritmer bestående av flere lag som ser etter spesifikke egenskaper i bildene. Dyplæringsalgoritmer er en mer avansert utgave av maskinlæring, med flere lag, som kan utføre mer kompliserte dataanalyser (26).

Dyplæringsalgoritmer, og spesielt konvolusjonelle nevrale nettverk (KNN), har på kort tid blitt den fremste metoden for å analysere radiologiske bilder (28, 29). Et KNN består av multiple lag, blant annet et konvolusjons-lag, og er designet for automatisk opplæring for identifisering av objekter (30). De ulike lagene fokuserer på ulike egenskaper ved objektet. U-Net er en versjon av KNN, som er egnet for færre sett av treningsdata (31). Det gjør den spesielt egnet for algoritmer som trenes opp med manuelt segmenterte data.

Segmenteringsalgoritmer må «trenes opp», det vil si at den må prosessere eksempler på det den skal utføre. For at den skal kunne segmentere bilder, må eksempler på utførte segmenteringer prosesseres. I denne prosesseringen, analyserer de ulike lagene, som beskrevet i avsnittet over, de ulike aspektene i bildene. Algoritmen som utarbeides ved prosesseringen, er enkelt sagt er en oppskrift for hvordan segmenteringen skal utføres. Algoritmen kan ikke utføre bedre segmenteringer med det den er trent opp med. Variasjoner i materialet den skal segmentere, påvirker hvor mange datasett som er nødvendige for opptrening. Ved stor variasjon vil det kreves større antall og motsatt. Dette innebærer at for hver gang den trenes opp, øker presisjonen i teorien. Kriteriet for økt presisjon gjelder når bildene som algoritmen skal segmentere ligner de den er trent på. Hvis algoritmen skal segmentere et datasett med andre anatomiske trekk, vil dette kunne medføre nedsatt presisjon.

3. Problemstillinger

Hensikten med masterstudien var å utvikle en metode for segmentering av vCSF og sCSF på MR-bilder av barn fra 0 til 2 år. Metodeutviklingen foregikk ved å trene opp en algoritme basert på dyplæring, heretter kalt KI-algoritmen, for automatisk segmentering. Hensikten var å utvikle en KI-algoritme som kunne utføre nær helautomatiske segmenteringer av vCSF og sCSF. Manuell korrigering ble utført etter den automatiske segmenteringen. I tillegg ble den utviklede metoden undersøkt med hensyn til validitet og reliabilitet. Evalueringene vil bidra til en vitenskapelig forankret metode for volummåling av CSF.

Metodestudien har undersøkt følgende problemstillinger:

1) Samsvar mellom KI-algoritmen og gullstandard, som et mål på validitet.
2) Utviklingen av KI-algoritmen fra versjoner 1-3 til versjoner 4-5.
3) Inter-rater reliabilitet mellom Rater 1 og Rater 2, for en gruppe med «normal CSF» og en gruppe med «økt CSF»
4) Intra-rater reliabilitet for Rater 2 for «normal CSF»

4. Materiale og metode

4.1 Studiedesign

Masterstudien er en kvantitativ metodestudie med analyser av validitet og reliabilitet.

4.2 Manuell segmentering, SliceOmatic

For manuell segmentering og korrigering ble programmet SliceOmatic (Tomovision, Montreal, Canada) (32) benyttet. SliceOmatic er et semi-automatisk bildeanalyse-program for segmentering av MR eller CT bilder. Semi-automatisk segmentering innebærer å sette terskelverdier for bestemte gråtoner, men fordrer at bildene fremstiller en aktuell struktur i tilnærmet samme gråtone. Grunnet stor variasjon av gråtoner og artefakter i MR bilder, er de semi-automatiske prosessene i analyseprogrammet lite anvendelige, og blir en hovedsakelig manuell prosess.

Programmet genererer tall for volum, i ml og antall piksler som er segmentert. For 2D bilder estimeres vokselstørrelsen, som er piksel-størrelsen ganget med snittykkelsen. Teknisk informasjon er lagret i datasettene i såkalte DICOM-tags. Antall voksler SliceOmatic estimerer, er identisk med antall piksler i datasettene. Alle manuelle segmenteringer og korrigeringer ble utført i SliceOmatic.

Prosedyre for manuell segmentering i SliceOmatic er beskrevet i Vedlegg 2.

4.3 KI-algoritmen

KI-algoritmene ble utviklet med et dataprogram produsert i 2018 av Tomas Sakinis (Avdeling for radiologi, Rikshospitalet, OUS) og basert på arbeid fra 'Radiology Informatics Lab', ved Mayo klinikken, USA (33). KI-algoritmen ble laget med mål om å utføre nær helautomatiske segmenteringer av vCSF og sCSF.

Innføringen av KI reduserte tiden for segmentering per datasett fra 8-10 timer til under 30 minutter. Selve KI-algoritmen brukte mindre enn 30 sekunder på segmenteringen. Denne KI-segmenteringen omtales som prediksjon fordi den kan inneholde feilsegmentering i form av under- eller over-estimering. Manuell korrigering ble derfor utført i etterkant i SliceOmatic.

KI-algoritmen er basert på programmeringsspråket *Python* og dyplærings bibliotekene Tensorflow og Keras. Programmet er en U-Net versjon av konvolusjonelle nevrale nettverk.

4.4 Metodeutvikling

Metodeutviklingen foregikk i tidsrommet 2018-2021. I oppstartsfasen ble valg av den best egnede MR-sekvensen for segmentering testet ut i konsensus med medarbeidere. De to raterne (MR-radiografer) gjennomgikk opplæring i programmet SliceOmatic. I starten var all segmentering basert på manuelt arbeid i SliceOmatic og opplæring i anatomi, standardisering og utarbeiding av en fasit for manuell segmentering av vCSF og sCSF på T2 vektede bilder. Denne fasiten for manuell segmentering blir heretter benevnt som gullstandard.

To erfarne MR-radiografer utførte samtlige segmenteringer. For evaluering av inter- og intra-rater reliabilitet blir radiografene heretter referert til som Rater 1 og Rater 2.

Valg av MR sekvens

For utviklingen av en konsekvent metode, måtte det velges kun én type MR-sekvens, og i et bestemt plan. Ulike sekvenser med ulike vektinger, T1 og T2, ble testet. **T2 aksial** i 2D format ble valgt som segmenteringssekvens, da den viste seg å være den enkleste med henhold til kontraster og snittplan. I tillegg er dette en av de mest anvendte sekvensene, og inngår i de fleste MR-undersøkelser av hjernen. **T1 vektete sekvenser**, som fremstiller CSF med lavt (mørkt) signal, viste seg å være vanskelig å differensiere fra ben (kraniet). De koronale og sagittale planene var mest utfordrende når det kom til å skille mellom vCSF og sCSF.

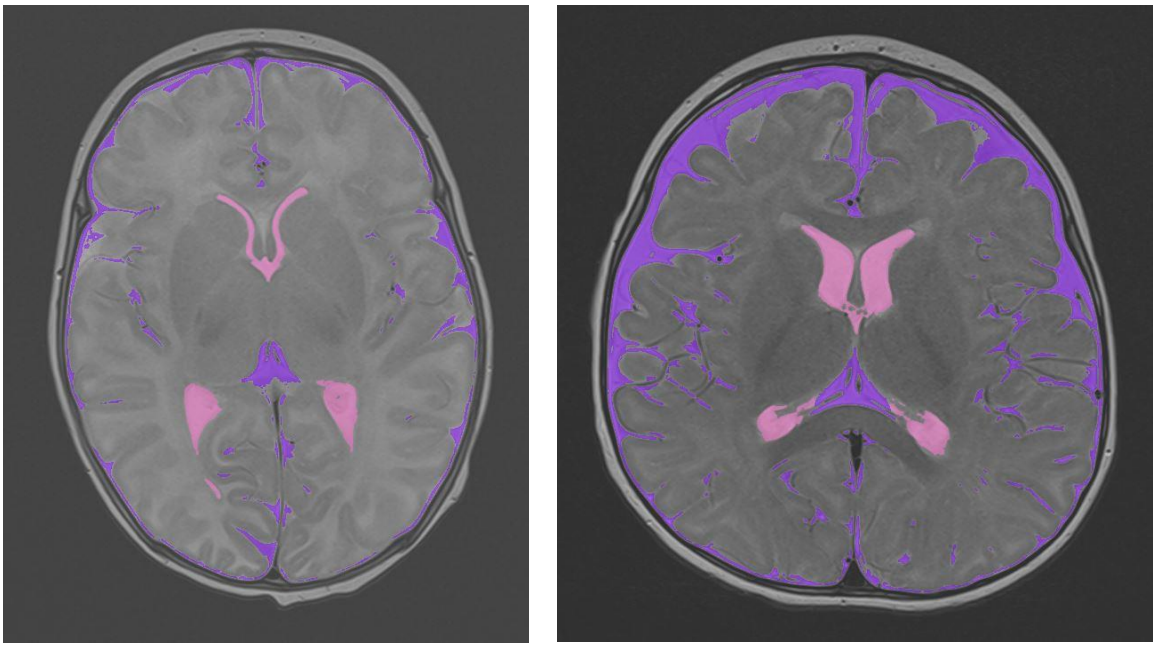
vCSF og sCSF

Med bakgrunn i tilstander som fører til økt CSF volum i subaraknoidalrommet, som beskrevet i bakgrunnskapittelet, skiller studien mellom CSF i ventriklene (vCSF) og CSF i subaraknoidalrommet (sCSF).

CSF befinner seg både rundt og inne i hjernen, og i tillegg rundt ryggmargen. I denne studien valgte vi å segmentere fra overgangen mellom ryggmargen og hjernen (foramen magnum). Årsaken til dette skyldes at det er langt færre undersøkelser som kombinerer hodet og rygg.

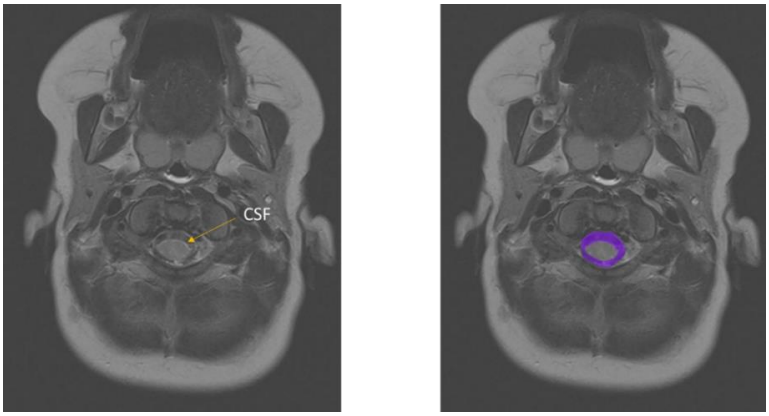
Gullstandard

I denne studien er gullstandard definisjonen på et datasett ideelt segmentert for vCSF og sCSF, som vist i Figur 7. Begrepet brukes om manuelt korrigerte datasett, utført etter automatisk segmentering.



Figur 7 **Gullstandard-segmentering av to hjerner,** vCSF markert i rosa, sCSF i lilla. Alderen for hjernen til venstre er 1 måned, til høyre er 6 måneder. Bildet til høyre viser lett markert sCSF volum, beskrevet i radiologisvar. Bildene tilhører studiens datasett

De to raterne ble veiledet av erfarne nevroradiologer i hjernens MR-anatomi. Avgrensningene av ventrikkelsystemet ble gjennomgått nøye, da det ikke alltid er noen klar grense (kontrast) mellom vCSF og sCSF. Det ble utarbeidet en felles forståelse for hvor detaljert CSF systemet skulle segmenteres. En prosedyre ble laget for hva som skulle inkluderes og ekskluderes, et eksempel er definisjon av første snitt som ble snittet under lillehjernen, på nivå med foramen magnum. Grundig anatomisk gjennomgang sikret at raterne gjenkjente feilregistrering av CSF ved strømnings-artefakter, der pikslene med CSF får signaltap og fremstår mørke, som vist i Figur 8.



Figur 8 Flow artefakt i CSF: signaltap i CSF som ligger rundt medulla (ryggmargen), på snitt ved foramen magnum. CSF fremstår mørk grå i et T2 vektet aksialt bilde, hvor den teoretisk skal fremstå lys. Bilde til høyre viser CSF segmentert i lilla. Bildene tilhører studiens datasett.

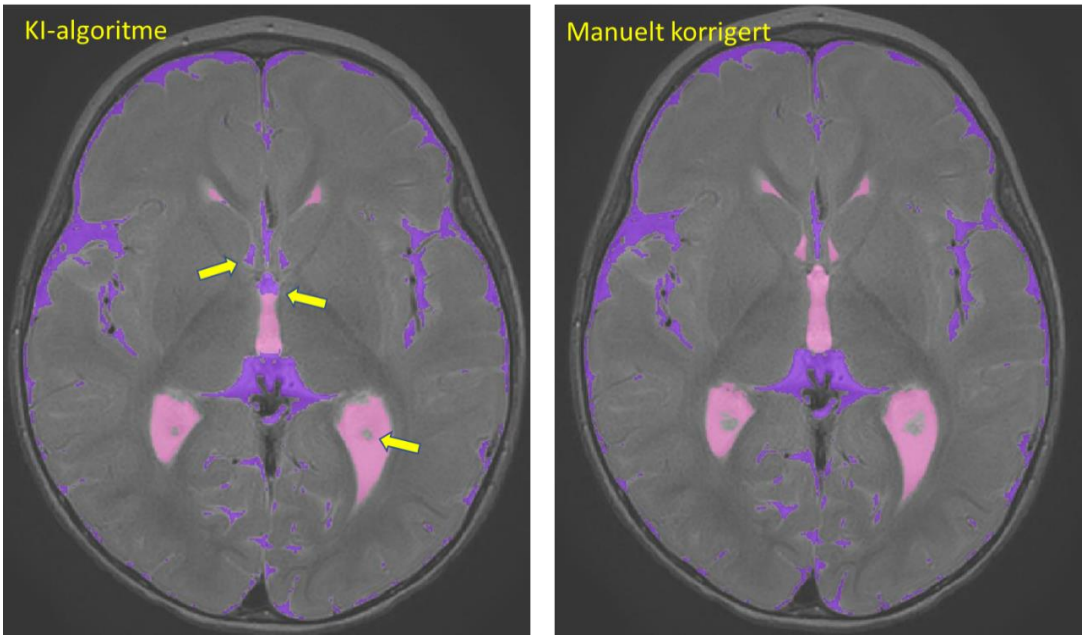
En medisinsk fysiker veiledet raterne om tilnærming til partiell volumeffekt i bildene. Blandingsvoksler som bestod av både hjernematerie og CSF, måtte segmenteres konsekvent. Tilnærmingsmetoden ble å inkludere 50 % av de pikslene med usikkerhet. Sannsynligheten for piksler med CSF ble også vurdert ved å studere snitt over og under det aktuelle snittet.

Fra manuell til automatisk segmentering

Datasettene med 20-40 bilder tok mellom 8 og 10 timer å segmentere manuelt. Det ble derfor innsett allerede i startsfasen, at den manuelle metoden måtte automatiseres. Siden målet var å samle inn referanseverdier for CSF, anslo vi at materialet måtte være på over 1000. Etablerte automatiske segmenteringsprogrammer ble undersøkt for mulig implementering. Utfordringen med de automatiske programmene, var at de krevde spesifikke sekvenser som 3D opptak med 1 mm oppløsning, som for eksempel 'Freesurfer', et av de mest anvendte segmenteringsprogrammer (34).

En radiolog og datautvikler i OUS hadde i dette tidsrommet utviklet en segmenteringsalgoritme for CT-bilder, og ideen ble overført til CSF studien. En egen algoritme ble utviklet for vCSF og sCSF segmentering på MR-bilder og implementert i studien. En egenutviklet algoritme hadde det fortrinn at den kunne trenes opp med egne segmenterte datasett og derfor designet for varierte T2 sekvenser i 2D format, tilpasset barnehjerner, og egnet for å differensiere mellom vCSF og sCSF.

KI-algoritmens første versjon ble trent opp med fem manuelt segmenterte datasett. De fem første ble gjennomgått i minste detalj, og ble definisjonen på hvordan gullstandard skulle se ut. I etterkant har hver ny KI-algoritme versjon blitt trent opp med 5-10 nye (gullstandard) datasett. For å oppnå presise segmenteringer på nivå med gullstandard, var den avhengige av flere treningsdata, segmenterte datasett. Det ble observert en gradvis, men varierende, forbedring for hver ny KI-algoritme versjon. Underestimering og feil markering av vCSF og sCSF var de største utfordringene, som vist i Figur 9. Feilsegmenteringer utført av KI-algoritmen krevde at manuell korrigering måtte utføres i etterkant, før datasettet ble godkjent, og kunne kvantifisere et volum. Opptreningsprosessen for hver ny versjon, der segmenterte bilder ble prosessert av dyplæringsprogrammet, tok ca. et døgn.



Figur 9 KI-algoritme segmentert og manuelt korrigert datasett. Bildet til venstre viser KI-algoritme med feilsegmentering. KI-algoritmen har segmentert sCSF i sideventriklene og tredje ventrikkel (gule piler). vCSF skal markeres i rosa og sCSF i lilla. Dette er rettet i bildet til høyre som viser manuell korrigering. KI-algoritmen har også overestimert vCSF i choroid plexus i sideventriklenes bakhorn, som er rettet opp til gullstandard-nivå i det manuelt korrigerte. Manuelt korrigert bilde viser i tillegg mer utfyllende sCSF i fissura Sylvi. Bildet tilhører studiens datasett.

Fremgangsmåte fra uthenting av datasett til ferdig volum

Fremgangsmåten per i dag, er som følger:

- Uthenting og avidentifisering av data fra radiologisk bildearkiv (PACS)
- Overføring av datasett til en forsknings PC
- Automatisk segmentering av datasettene i siste versjon av KI-algoritmen
- Manuell korrigering av automatisk segmenterte datasett i SliceOmatic
- Kalkulering av volum av manuell korrigerte i SliceOmatic

4.5 Evaluering

KI-algoritmen versus gullstandard og utvikling fra versjon 1-3 til 4-5

Samsvar mellom KI-algoritmen og gullstandard ble undersøkt, som et mål algoritmens validitet. For å undersøke utviklingen av KI-algoritmen ble det inkludert 20 datasett: 10 fra versjon 1-3, og 10 fra versjon 4-5, for å påvise eventuell økt presisjon fra de første versjonene til de to siste versjonene.

I begge grupper ble datasett segmentert av KI-algoritmen sammenlignet med 20 gullstandard. Gullstandard var først segmentert av algoritmen, deretter manuelt korrigert. Hvert datasett inneholdt to variabler: vCSF og sCSF.

Følgende variabler ble sammenlignet:

- KI-algoritme versjon **1-3** vs. gullstandard, **vCSF**
- KI-algoritme versjon **1-3** vs. gullstandard, **sCSF**
- KI-algoritme versjon **4-5** vs. gullstandard, **vCSF**
- KI-algoritme versjon **4-5** vs. gullstandard, **sCSF**

Tabell 1 Oversikt over KI-algoritme versjoner 1-5. KI-algoritme versjoner med antall datasett inkludert i gjeldende versjon og antall datasett som versjonen er trent opp med.

KI-algoritme versjon	Antall datasett inkludert fra versjonen	Antall datasett versjonen er opptrent med
1	4	5
2	3	10
3	3	15
4	5	25
5	5	35

Inter-rater reliabilitet

Inter-rater reliabilitet ble undersøkt for to grupper, en med normalt volum av CSF og en med økt volum av CSF. Utvalget ble gjort på bakgrunn av radiologisk beskrivelse av MR-undersøkelsen. Det ble valgt ut 10 datasett i «normal CSF» og 10 datasett i «økt CSF». Gruppen «økt CSF» inkluderte tilstander som benign ekstern hydrocephalus (BEH) og hydrocephalus. Hvert datasett inneholdt to variabler: vCSF og sCSF.

Datasettene ble segmentert av KI-algoritmen først (versjon 4 og 5 ble benyttet), og deretter manuelt korrigert, av henholdsvis Rater 1 og Rater 2.

Hensikten med utvalget «økt CSF», var å undersøke om økt volum ga lik reliabilitet som normalt volum. Raterne hadde mindre erfaring med segmentering av økt CSF volum. Siden bildene med økt CSF volum skilte seg visuelt ut fra normale CSF volum, var det ikke hensiktsmessig å blinde raterne i forhold til hvilken gruppe de segmenterte. KI-algoritmen var

i hovedsak trent opp med hjerner med normale volum. Det ble antatt redusert presisjon av KI-algoritmens segmentering ved økt volum, som ville føre til mer manuell korrigering.

Følgende variabler ble sammenlignet:

- Rater 1 vs. Rater 2, **vCSF** «**normal** CSF»
- Rater 1 vs. Rater 2, **sCSF** «**normal** CSF»
- Rater 1 vs. Rater 2, **vCSF** «**økt** CSF»
- Rater 1 vs. Rater 2, **sCSF** «**økt** CSF»

Intra-rater reliabilitet

Intra-rater reliabilitet ble undersøkt med 10 datasett, identisk med gruppen «normal CSF» fra inter-rater reliabilitet. Hensikten med identisk gruppe var for sammenligning med resultatene fra inter-rater reliabilitet. Hvert datasett inneholdt to variabler: vCSF og sCSF.

Datasettene ble segmentert av KI-algoritmen først (versjon 4 og 5 ble benyttet), og deretter manuelt korrigert av Rater 2. Det var 4 måneder mellom måletidspunktene.

Følgende variabler ble sammenlignet:

- Tidspunkt 1 vs. tidspunkt 2, **vCSF**
- Tidspunkt 1 vs. tidspunkt 2, **sCSF**

4.6 Materiale

Materialet i studien bestod av MR-undersøkelser av hodet fra barn mellom 0 og 2 år. Datasett ble hentet ut fra undersøkelsene som var tatt i klinisk sammenheng, og ikke primært til forskning. Alle bildene ble tatt ved OUS, i perioden 2011 til 2017. Datasettene ble hentet ut i perioden 2019 til 2020.

På inkluderingstidspunktet hadde studien samlet inn 70 datasett, og av disse ble det gjort et utvalg på 40 for evalueringene. Datasettene var hentet ut fra radiologisk bildearkiv i OUS, og avidentifisert. Utvalgene var i tilfeldig alder mellom 0 og 2 år. De ble overført til egen forsknings PC, avkoblet fra internett. Materialet har generert kontinuerlige variabler for volum, ml.

Inklusjonskriterier

- MR-undersøkelsen måtte inneholde en aksial T2 vektet sekvens og fullstendig hjernedekning fra skallebasis til øvre del av skallen
- Undersøkelsen måtte ha negative radiologifunn (være uten patologi), unntatt gruppen med «økt CSF»
- For gruppen «økt CSF»: radiologibeskrivelse av undersøkelsen måtte inneholde påvist økt CSF volum

Eksklusjonskriterier

- Bevegelsesartefakter i bildene, noe som kompliserer segmentering
- Lav oppløsning i bildene, piksler større enn 4 × 4 mm
- Radiologirapport med beskrevet patologi, unntatt gruppen med «økt CSF»

Tekniske parametre for datasettene

Et datasett består av en 2D aksial T2 sekvens med 20-40 snitt av en hjerne. Undersøkelsene er fra fem forskjellige skannere, og tre ulike leverandører, Phillips, GE og Siemens. Magnetstyrken var 1,5 eller 3 Tesla. Snittykkelsen i bildene var 3 eller 4 mm. Oppløsningen, piksel-størrelsen, varierte fra 2×2 mm til 3,9×3,9 mm. For utvalget tilsvarte 1 ml CSF fra 1042 til 4875 voksler.

4.7 Statistiske analyser

Dataene har blitt prosessert av Excel (Office 365) og SPSS (Statistical Package of Social Science) versjon 26/27. Volumdata er generert fra SliceOmatic (Tomovision, Quebec, Canada). Dice koeffisient er generert fra egen algoritme utviklet av radiolog Tomas Sakinis.

Vurderinger av normaldistribusjon av dataene ble gjort ved fremstilling i histogrammer og QQ plott. Differansevolum mellom variablene som ble sammenlignet, ble brukt som måleenhet, og viste varierende grad av normaldistribusjon. For analysene ble det tross variasjoner valgt å betrakte dataene som normalfordelte. Størrelsen av utvalg har en signifikant effekt på distribusjonen, og et lite utvalg gir sjelden normaldistribusjon (35). I denne studien var det 10 datasett per variabel, dette diskuteres videre i diskusjonskapittelet.

Matematisk evaluering - Dice koeffisient (piksel-overlapp)

Dice koeffisient, også kjent som Dice Similarity Coefficient (DSC), er en av de mest brukte metodene for validering av automatiske segmenteringsalgoritmer for bilder (9, 36).
Dice koeffisienten måler den faktiske overlappingen av piksler mellom to bildesett og ble regnet ut av en algoritme. To segmenterte MR-datasett ble sammenlignet av en algoritme for kalkulering av Dice koeffisient.

Formelen for Dice koeffisienten er: **2 (A∩B)/(A+B)**

Der A∩B tilsvarer piksel-overlapp mellom to datasett (A og B) som sammenlignes.
A+B er antall piksler summert for A og B (20). Dice koeffisienten blir et tall mellom 0 og 1, hvor 0 er ingen overlapp og 1 er fullkommen overlapp.

Beregningen av overlapp av piksler mellom datasettene som sammenlignes, er en matematisk evaluering av KI-algoritmen. Dette vil supplere den kliniske evalueringen, der samsvaret i volum (ml) sammenlignes.

A

1	1	1	1
1	0	0	1
1	0	0	1
1	1	1	1

B

0	1	1	1
1	1	1	1
1	1	0	1
1	1	1	1

Piksel-overlapp for A og B = 12 piksler

2 (12) / (16+16) = 0,75 (Dice koeffisient)

Figur 10 Eksempel på utregning av Dice koeffisient. Svarte tall tilsvarer piksel-overlapp (12 piksler) og røde tall tilsvarer ikke-overlapp (4 piksler). Matrisen til A og B består av 16 piksler hver.

Dice koeffisient ble rapportert for piksel-overlapp mellom variablene KI-algoritme vs. gullstandard, og for inter- og intra-rater reliabilitet. Dice koeffisient for hver enkelt KI-algoritme versjon vs. gullstandard, er i tillegg presentert i spredningsplott.

Klinisk evaluering - samsvar av volum

For reliabiliteten ble intraklasse korrelasjonskoeffisient (ICC) valgt, med følgende variant: *Two-way mixed effects, absolute agreement, single rater/measurement*. For validitet og inter-rater reliabilitet ble *Average measurement* valgt, for intra-rater reliabilitet, *single measurement*.

‘Two-way mixed effects’ brukes når de valgte raterne er de eneste raterne av interesse, som dermed betyr at resultat ikke kan generaliseres til andre ratere. Et absolutt samsvar, absolute agreement, velges når dette er viktigere enn konsistent samsvar.

ICC måler samsvar mellom ulike datasett av samme klasse, det vil si samme varians og metriske målestokk, for kontinuerlige variabler. Det gir et resultat med et tall mellom 0-1, der 1 er høyest samsvar og forutsetter normaldistribusjon (37). ICC ble rapportert for samsvar mellom KI-algoritme vs. gullstandard, og for inter- og intra-rater reliabilitet.

Tolkning av reliabilitetskoeffisienten kan deles opp etter Munro’s kategorisering av styrken til en korrelasjonskoeffisient (38), som vist i Tabell 2. Denne oppdelingen er ikke absolutt, og avhengig av hva som sammenlignes og kontekst (39).

Tabell 2 Munro’s kategorisering av korrelasjonskoeffisient:

Reliabilitetskoeffisient	Enighet
,00-,25	Liten, hvis noe, korrelasjon
,26-,49	Lav korrelasjon
,50-,69	Moderat korrelasjon
,70-,89	Høy korrelasjon
,90-1,00	Svært høy korrelasjon

For å sammenligne ulike metoder er det anbefalt å se på differansene mellom datasett i tillegg til enighet (40). I 1983 introduserte Bland og Altman en alternativ metode som siktet til å fremstille enighet mellom to metoder på en mer nøyaktig måte enn standarden med korrelasjonsanalyser (41). Bland-Altman plottet fremstiller gjennomsnittsdifferanse mot gjennomsnittsmåling, mellom to datasett, og et intervall av enighet, der 95 % av differansene befinner seg. Fordelen med slik analyse er å synliggjøre skjevheter. Analysen sier ikke noe om hvorvidt differansene er akseptable, det må tolkes opp mot klinisk relevans.

I en publikasjon (1998) med mål om å lage en statistisk guide for reliabilitetsstudier, konkluderte Rankin og Stokes med at ICC og Bland-Altman plott passer godt for reliabilitetsstudier, men de gir de ikke adekvat informasjon hver for seg, og anbefales derfor brukt sammen (42).

4.8 Etiske betraktninger

Masterprosjektet er en del av en overordnet studie, som er godkjent av REK, vedtak 2018/2510, (vedlegg 1) og personvernombudet (PVO) ved OUS. Det ble ikke innhentet ny informasjon, studien medførte ingen risiko for de inkluderte. Resultatene fra studien vil ikke kunne identifisere enkeltindivider.

5. Resultater

5.1 Validitet: KI-algoritmen versus gullstandard, for versjoner 1-3 og 4-5

KI-algoritme versjon 1-3, underestimerte vCSF sammenlignet med gullstandard med 10 %, og sCSF med 3 %, mens underestimeringen i versjon 4-5 var 0,7 % for vCSF og 0,4 % for sCSF. Økningen i volum fra versjon 1-3 til 4-5, var betinget i alderen til barna, gjennomsnittsalder og range vist i Tabell 3.

Tabell 1 Gjennomsnittsvolum målt ved KI-algoritme og gullstandard, i versjon 1-3 og 4-5.
Volum av vCSF og sCSF med gjennomsnitt og standard målefeil, hos 10 barn i KI-algoritme versjon 1-3 og 10 barn i KI-algoritme versjon 4-5. Tilfeldig utvalgt alder mellom 0-2 år, gjennomsnittsalder er rapportert under tabell. Datasettene i KI-algortime versjon 1-3 og 4-5 er ikke identiske.

	vCSF Gjennomsnittsvolum ± SEM **ml**	**sCSF** Gjennomsnittsvolum ± SEM **ml**
KI-algoritme versjon 1-3, n =10*		
KI-algoritme	8,9 ± 1,3	63,1 ± 10,6
Gullstandard	9,9 ± 1,4	64,9 ± 10,0
KI-algoritme versjon 4-5, n=10**		
KI-algoritme	14,8 ± 2,3	84,5 ± 8,5
Gullstandard	14,9 ± 2,3	84,2 ± 8,6

**Gjennomsnittsalder(range) i uker: 44 (1-97)*
***Gjennomsnittsalder (range) i uker: 61 (31-104)*

SEM: standard målefeil
vCSF: ventrikkel cerebrospinalvæske
sCSF: subaraknoidal cerebrospinalvæske
KI: Kunstig intelligens
Gullstandard: manuell korrigering
ml: milliliter

Presisjonen av piksel-overlapp økte fra versjon 1-3 til 4-5, Dice koeffisienten gikk fra 0,83 til 0,97 for vCSF og fra 0,83 til 0,99 for sCSF. ICC, basert på samsvar av volum, viste resultater ≥ 0,985 for samtlige versjoner av KI-algoritmen, men høyest for versjon 4-5. Konfidensintervallet, ICC, og gjennomsnittsendring av antall piksler ble redusert for vCSF i versjon 4-5 sammenlignet med versjon 1-3. Gjennomsnittsdifferansene bekreftet positiv utvikling for KI-algoritmens versjon 4-5 med lavere gjennomsnittsdifferanser og intervall for likhetsgrader (95%).

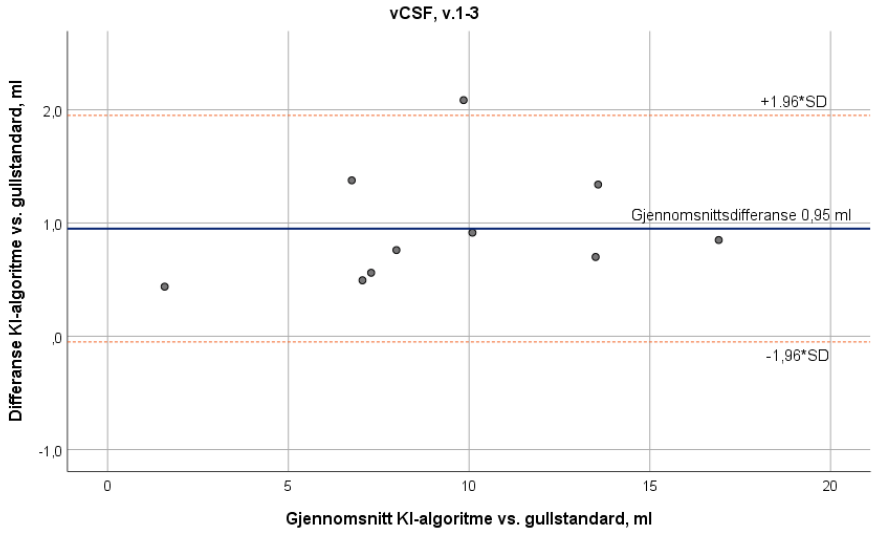
Tabell 2 Validitet for KI-algoritmen. Validitet vist som samsvar mellom KI-algoritmen og gullstandard med Dice koeffisient, ICC, gjennomsnittsdifferanser og gjennomsnittsendring av piksler. Den ene gruppen er KI-algoritme versjon 1-3 og den andre er KI-algoritme versjon 4-5.

	Dice koeffisient	ICC (95 % Konfidensintervall)	Gjennomsnittsdifferanse (95 % LoA) ml	Endring av antall piksler*: gjennomsnitt (range)
KI-algoritme v. 1-3, n =10				
KI vs. Gullstandard, vCSF	0,83	0,985 (0,343-0,998)	0,9 (-0,0-1,9)	2521 (756-9982)
KI vs. Gullstandard, sCSF	0,83	0,987 (0,950-0,997)	1,8 (-13,1-16,7)	8172 (-15459-60778)
KI-algoritme v. 4-5, n=10				
KI vs. Gullstandard, vCSF	0,97	1,000 (0,998-1,000)	0,2 (-0,2-0,5)	355 (-231-1298)
KI vs. Gullstandard, sCSF	0,99	1,000 (0,999-1,000)	0,3 (-1,1-0,5)	-805 (-3115-86)

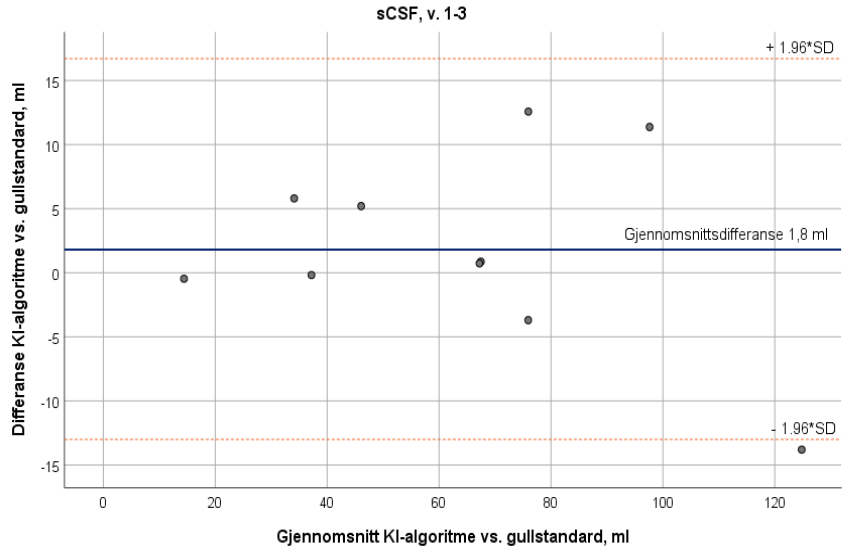
**Grunnet datasettene har ulik oppløsning og den totale mengden piksler varierer, er dette kun inkludert som tilleggsinformasjon til resultatene og kan ikke sammenlignes på tvers. Antall piksler er identisk med antall voksler.*

vCSF: ventrikkel cerebrospinalvæske
sCSF: subaraknoidal cerebrospinalvæske
KI: Kunstig intelligens
Gullstandard: manuell korrigering
ml: milliliter
LoA: Limits of Agreement

Tendens til større differanser ved økt volum, ble observert i Bland-Altman plott for sCSF i versjon 1-3, vist i Figur 11b. I versjon 4-5, ble ikke differansene påvirket av volumet, og intervallet der 95 % vil ligge, ble redusert, som vist i Figur 12a og b.

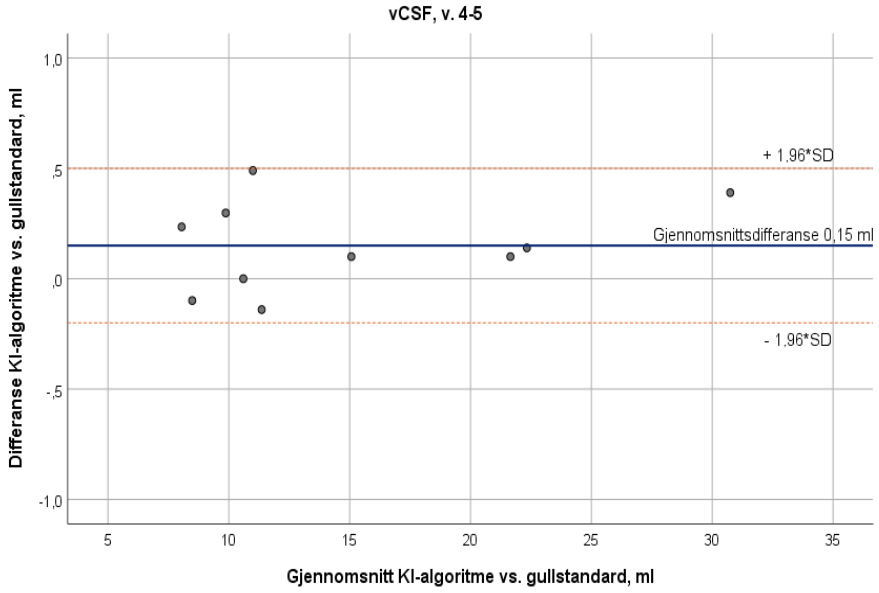


Figur 11a

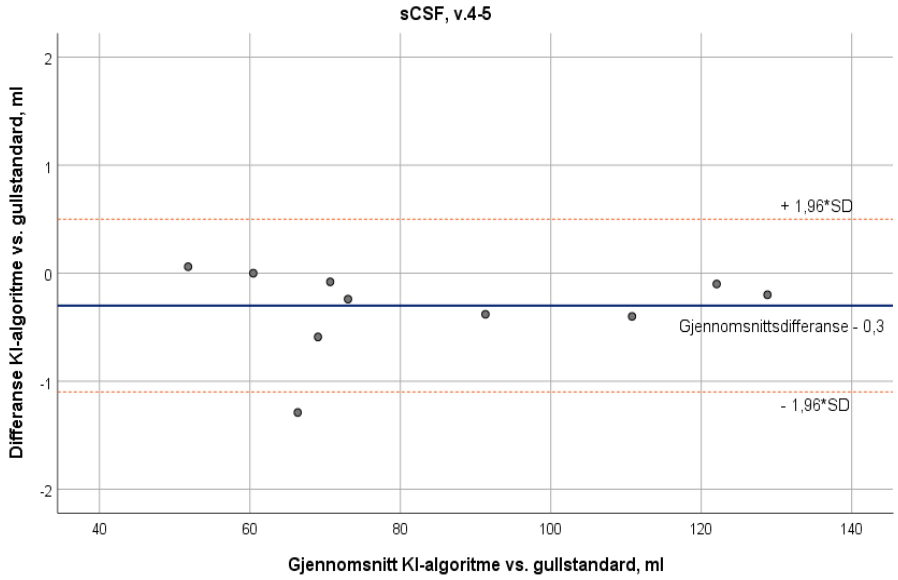


Figur 11b

Figur 11 Bland-Altman plott vCSF (11a) og sCSF KI (11b) KI-algoritme vs. gullstandard versjon 1-3, n=10: Gjennomsnittsdifferanse mellom KI-algoritme 1-3 og gullstandard er vist langs y-aksen (gullstandard - KI-algoritme) og gjennomsnitt mellom KI-algoritme og gullstandard langs x-aksen (gullstandard + KI-algoritme/2).
LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD



Figur 12a



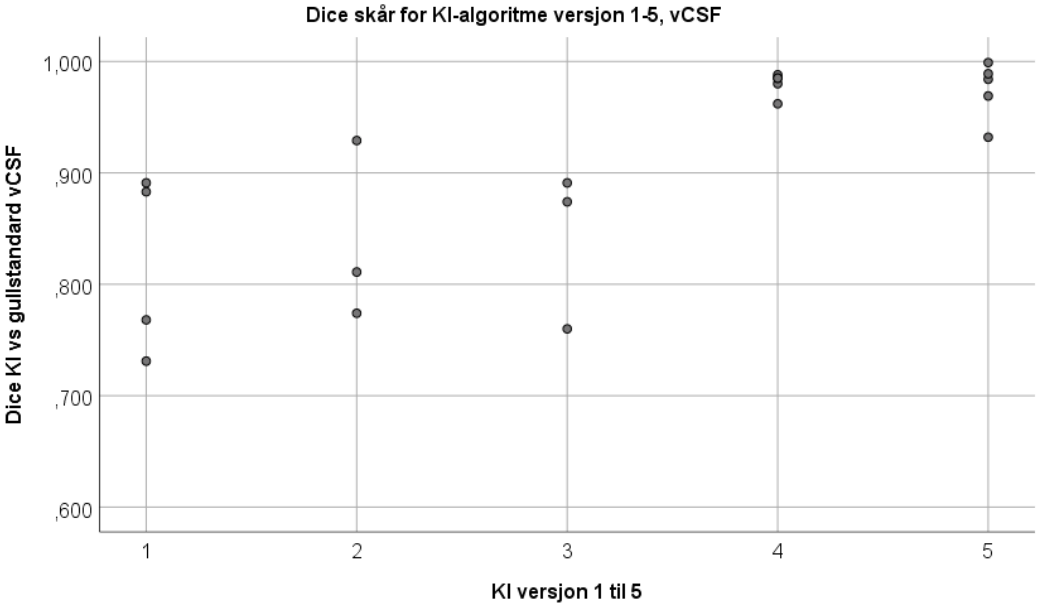
Figur 12b

Figur 12 Bland-Altman plott vCSF (12a) og sCSF (12b) KI-algoritme vs. gullstandard versjon 4-5, n=10: Gjennomsnittsdifferanse mellom KI-algoritme 4-5 og gullstandard er vist langs y-aksen (gullstandard - KI-algoritme) og gjennomsnitt mellom KI-algoritme og gullstandard langs x-aksen (gullstandard + KI-algoritme/2).
LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD.

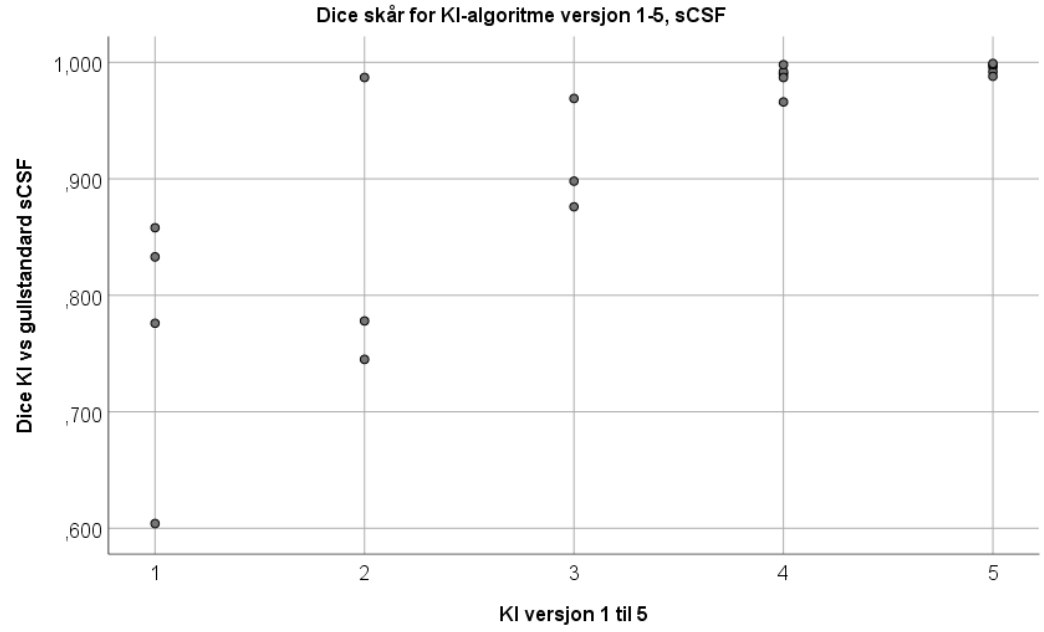
Piksel-overlapp for KI-algoritme versjoner 1-5

Piksel-overlapp mellom KI-algoritmen og gullstandard for hver versjon fra 1 til 5, er fremstilt i spredningsplott i Figur 13a og b.

Stor spredning av Dice koeffisient for samtlige av de tre første versjonene for vCSF og sCSF ble bekreftet, som vist i Figur 13a og b. Minst spredning ble observert for sCSF i versjon 5. I versjon 2, for sCSF, hadde en observasjon like høy Dice koeffisient som versjon 4-5, vist i Figur 13b. Kun en observasjon hadde Dice koeffisient under akseptabel verdi som er 0,7, vist i Figur 13b for KI-algoritme versjon 1.



Figur 13a



Figur 13b

Figur 13 Spredningsplott for Dice-koeffisient for hver KI-algoritme versjon, vCSF (13a) og sCSF (13b), n=20: Piksel-overlapp for KI-algoritme vs. gullstandard for versjon 1-5. Antall data per KI-algoritme versjon er for versjon 1: n= 4, versjon 2: n=3, versjon 3: n= 3, versjon 4 n=5, og versjon 5: n=5.

5.2 Inter-rater reliabilitet for «normal CSF» og «økt CSF»

Det var høy overenstemmelse mellom Rater 1 og Rater 2, som vist i Tabell 5. For vCSF i gruppen «normal CSF» var det absolutt samsvar mellom raterne, for sCSF, er det en differanse på 0.4%. I gruppen «økt CSF» var det en differanse mellom raterne på 1% for vCSF, og 2 % for sCSF.

Gjennomsnittsvolum var nær det dobbelte for «økt CSF» sammenlignet med «normal CSF».

Tabell 3 Gjennomsnittsvolum målt av Rater 1 og Rater 2, ved gruppen «normal CSF» og «økt CSF»: Segmentert volum for vCSF og sCSF hos 10 barn med normalt CSF volum og 10 barn med økt CSF volum.

	vCSF Gjennomsnittsvolum ± SEM **ml**	**sCSF** Gjennomsnittsvolum ± SEM **ml**
«Normal CSF», n=10*		
Rater 1	15,3 ± 2,3	94,2 ± 7,0
Rater 2	15,3 ± 2,3	94,6 ± 6,9
«Økt CSF», n=10**		
Rater 1	34,3 ± 3,7	180,8 ± 13,6
Rater 2	34,8 ± 3,7	185,3 ± 14,1

*****Gjennomsnittsalder (range) antall uker: 86 (28-76)
**Gjennomsnittsalder (range) antall uker: 40 (28-48)

vCSF: ventrikkel CSF
sCSF: subaraknoidal CSF
ml: milliliter
SEM: standard målefeil

Dice koeffisienten var tilnærmet lik for begge gruppene, men noe høyere for sCSF i gruppen «normal CSF». ICC viste høyt samsvar for samtlige med resultater ≥ 0,997, men større konfidensintervall for sCSF i gruppen «økt CSF». Gjennomsnittsdifferansene og LoA intervall var noe større i «økt CSF», mest tydelig for sCSF, der alle resultatene viste noe lavere samsvar.

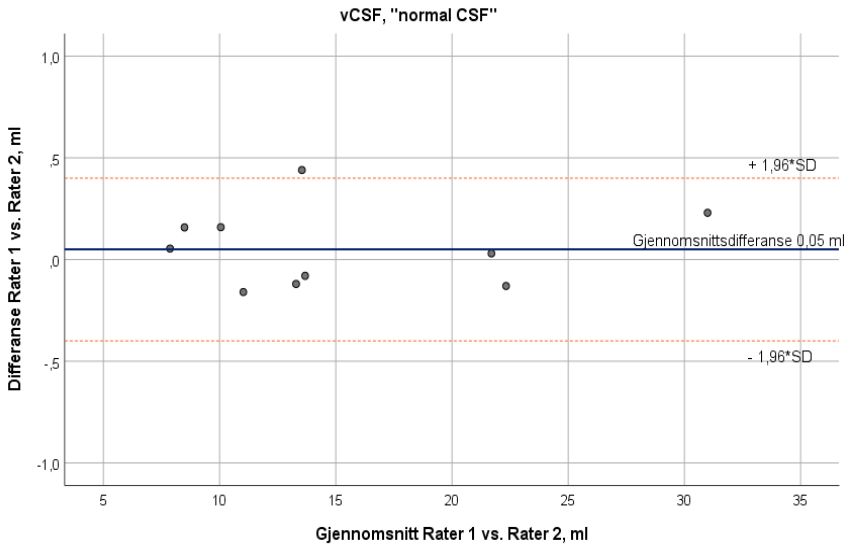
Tabell 4 Inter-rater reliabilitet for gruppene «normal CSF» og «økt CSF»:
Samsvar mellom Rater 1 og Rater 2 vist ved Dice koeffisient basert på piksel-overlapp, ICC, gjennomsnittsdifferanse og gjennomsnittsendring av antall piksler.

	Dice koeffisient	ICC (95 % Konfidensintervall)	Gjennomsnittsdifferanse (95 % LoA) ml	Endring av antall piksler*: gjennomsnitt (range)
«Normal CSF», n=10				
Rater 1 vs. Rater 2, vCSF	0,97	1,000 (0,999-1,000)	0,1 (-0,3-0,4)	148 (-279-713)
Rater 1 vs. Rater 2, sCSF	0,98	1,000 (0,998-1,000)	0,4 (-1,1-2,0)	914 (-1964-4718)
«Økt CSF», n=10				
Rater 1 vs. Rater 2, vCSF	0,97	0,999 (0,989-1,000)	0,5 (-0,4-1,5)	753 (-134-2667)
Rater 1 vs. Rater 2, sCSF	0,96	0,997 (0,693-1,000)	4,4 (0,3-8,6)	6332 (2617-11387)

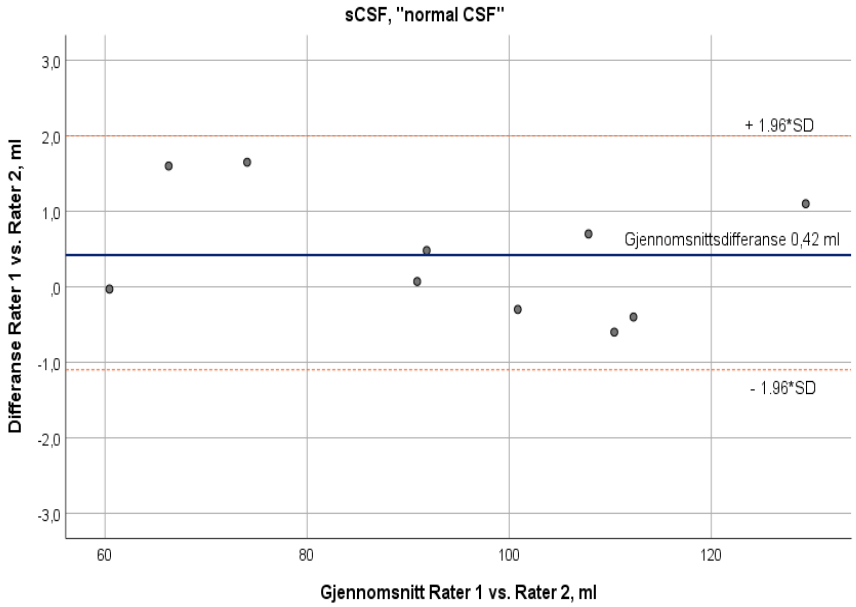
**Grunnet datasettene har ulik oppløsning og den totale mengden piksler varierer, er dette kun inkludert som tilleggsinformasjon til resultatene og kan ikke sammenlignes på tvers. Antall piksler er identisk med antall voksler.*

vCSF: ventrikkel cerebrospinalvæske
sCSF: subaraknoidal cerebrospinalvæske
LoA: Limits of Agreement
ml: milliliter

For gruppen «normal CSF» viste Bland-Altman plottet ingen systematiske feil eller uteliggere. Differansene mellom raterne økte ikke ved økt volum i «normal CSF», vist i Figur 14a og b. For vCSF i «økt CSF» var det tendens til økt differanse, ved økt volum, vist i Figur 15a. Differansene for sCSF i «økt CSF» var alle over 0, altså Rater 2 segmenterte mer enn Rater 1 på samtlige, vist i Figur 15b. To uteliggere ble sett for sCSF med en gjennomsnittsdifferanse på $\geq$ 7 ml, som vist i Figur 15b.

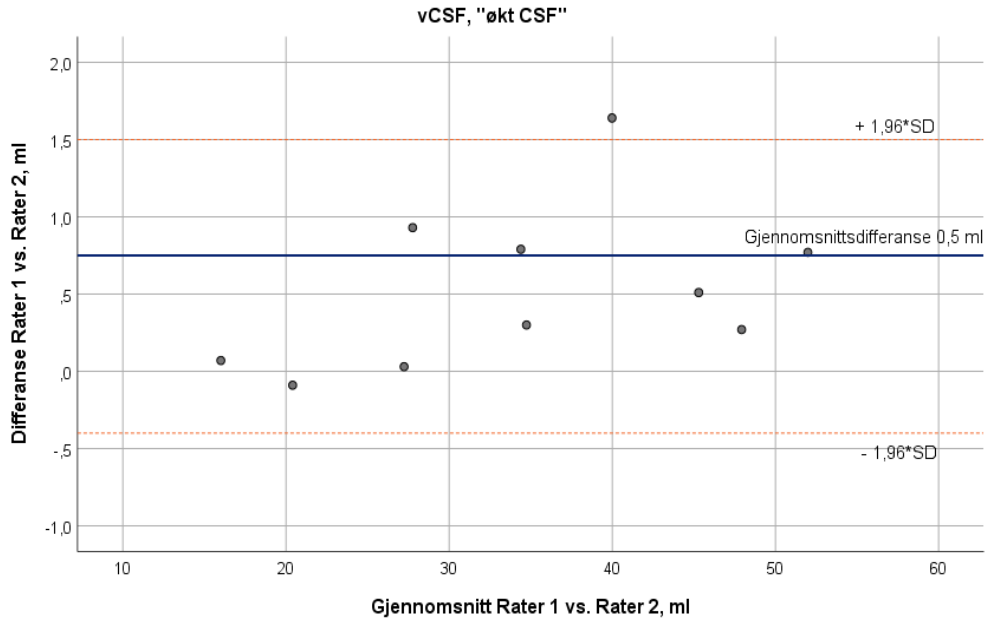


Figur 14a

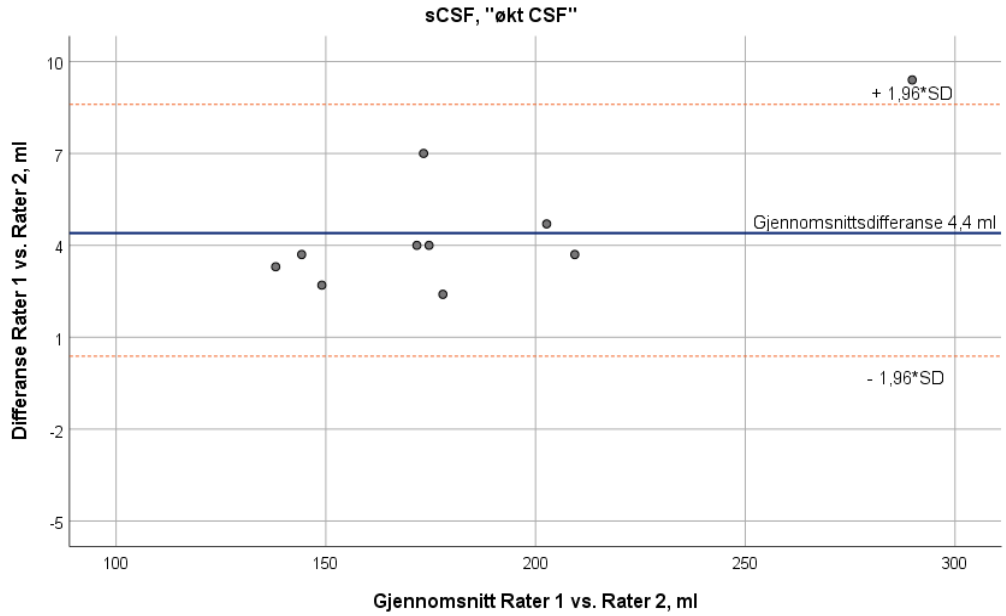


Figur 14b

Figur 14 Bland-Altman Plott for inter-rater reliabilitet, vCSF (14a) og sCSF (14b), «normal CSF», n=10: Gjennomsnittsdifferanse mellom Rater 1 og Rater 2 er vist langs y-aksen (Rater 2 – Rater 1) og gjennomsnitt mellom raterne langs x-aksen
(Rater 1 + Rater 2/2). LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD.



Figur 15a



Figur 15b

Figur 15 Bland-Altman Plott for inter-rater reliabilitet, vCSF (15a) og sCSF (15b), «økt CSF», n=10: Gjennomsnittsdifferanse mellom Rater 1 og Rater 2 er vist langs y-aksen (Rater 2 – Rater 1) og gjennomsnitt mellom raterne langs x-aksen (Rater 1 + Rater 2/2). LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD.

5.3. Intra-rater reliabilitet

Det var minimal forskjell mellom måling 1 og måling 2 for vCSF og sCSF. Gjennomsnittsvolumet hadde en differanse på 0,6 % for vCSF og 0,9 % for sCSF, som vist i Tabell 7.

Tabell 5 Gjennomsnittsvolum for Rater 2 ved to måletidspunkt, n=10*:
vCSF og sCSF volum hos 10 barn, ved to måletidspunkt. Rater 2 har utført segmenteringene og det er fire måneder mellom tidspunktene.

	vCSF Gjennomsnittsvolum ± SEM **ml**	**sCSF** Gjennomsnittsvolum ± SEM **ml**
Måling 1	15,3 ± 2,3	94,64 ± 6,9
Måling 2	15,4 ± 2,3	95,51 ± 7,0

** Datasettene er identiske med gruppen «normal CSF» fra inter-rater reliabilitet.*

vCSF: ventrikkel CSF
sCSF: subaraknoidal CSF
ml: milliliter
SEM: standard målefeil

ICC og konfidensintervallet viste svært høyt samsvar, for både vCSF og sCSF. Gjennomsnittsdifferansene på 0,1 ml for vCSF og 0,9 ml for sCSF, bekreftet høyt samsvar. Flere piksler ble i gjennomsnitt endret for sCSF, enn tilsvarende datasett i inter-rater reliabilitet, men påvirket ikke Dice koeffisienten.

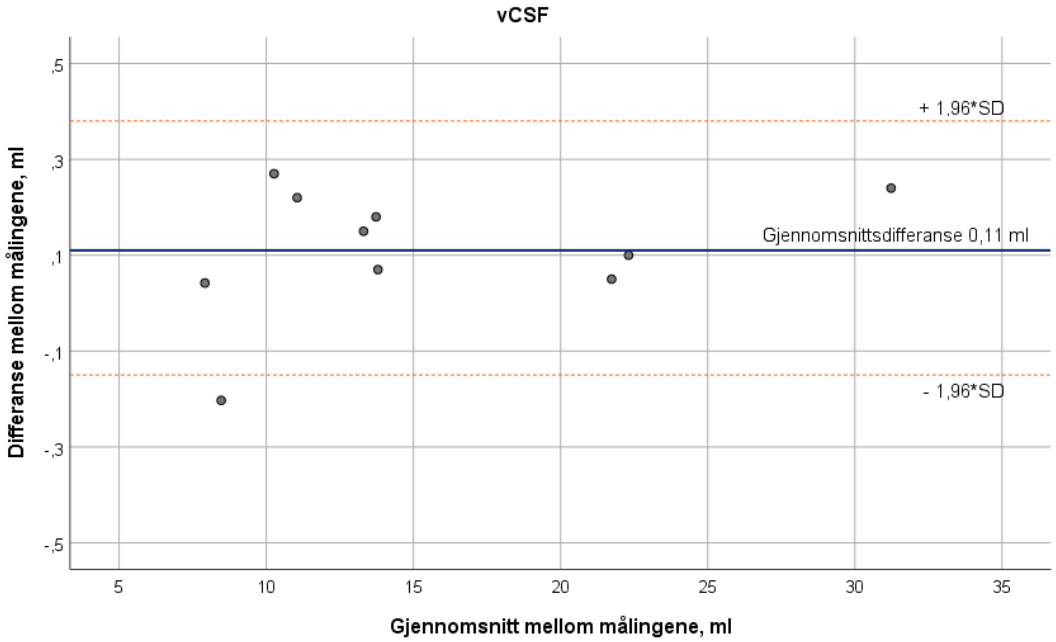
Tabell 6 Intra-rater reliabilitet for Rater 2: Samsvar mellom måletidspunkt vist ved Dice koeffisient (piksel-overlapp), ICC (volum), gjennomsnittsdifferanse (volum) og gjennomsnittsendring av antall piksler. Måleintervall er 4 måneder.

	Dice koeffisient	**ICC (95 % KI)**	**Gjennomsnittsdifferanse (95 % LoA) ml**	**Endring av antall piksler*: gjennomsnitt (range)**
Intra-rater, vCSF	0,97	1,000 (0,998-1,000)	0,1 (-0,2-0,4)	237 (-310-1175)
Intra-rater, sCSF	0,98	0,999 (0,969-1,000)	0,9 (-0,4-2,2)	1945 (270-8863)

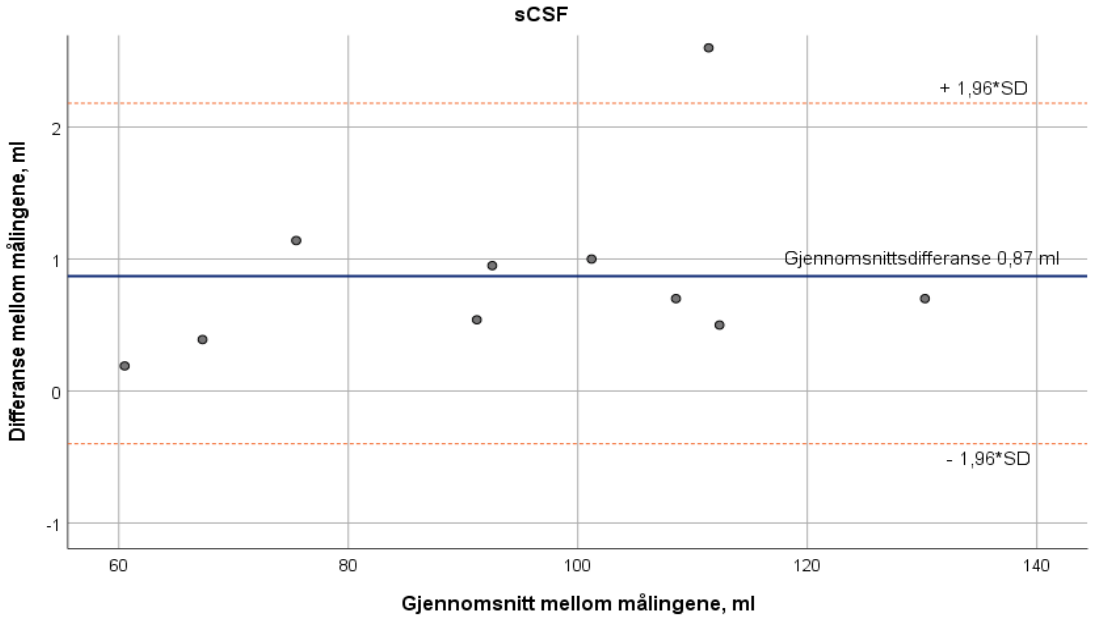
**Grunnet datasettene har ulik oppløsning og den totale mengden piksler varierer, er dette kun inkludert som tilleggsinformasjon til resultatene og kan ikke sammenlignes på tvers. Antall piksler er identisk med antall voksler.*

vCSF: ventrikkel cerebrospinalvæske
sCSF: subaraknoidal cerebrospinalvæske
LoA: Limits of Agreement
ml: milliliter

Bland-Altman plottene viste høyt samsvar og liten grad av spredning for vCSF og sCSF, med smalt intervall, der 95 % vil befinne seg, vist i Figur 16a og b. For sCSF var det liten spredning, 9/10 var nærme gjennomsnittsdifferansen på 0,87 ml, unntatt en differanse på over 2 ml, som vist i Figur 16b. For sCSF hadde alle gjennomsnittsdifferanser på > 0 ml, det ble målt større volum på det ene måletidspunktet.



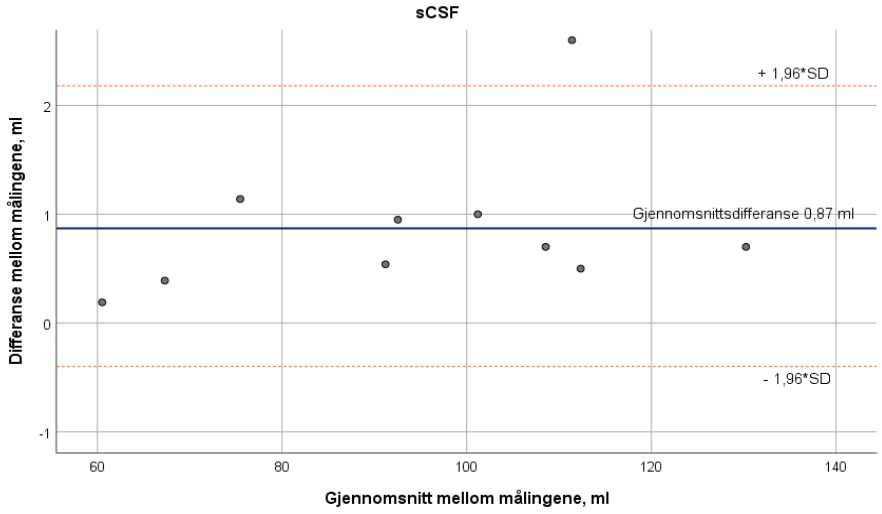
Figur 16a



Figur 16b

Figur 16 Bland-Altman Plott intra-rater reliabilitet for vCSF (16a) og sCSF (16b), n=10:
Gjennomsnittsdifferanse mellom to måletidspunkt er vist langs y-aksen (tidspunkt 2 – tidspunkt 1) og gjennomsnitt mellom måletidspunktene langs x-aksen (tidspunkt 1 + tidspunkt 2/2).
LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD

.



Figur 16b

Figur 16 Bland-Altman Plott intra-rater reliabilitet for vCSF (16a) og sCSF (16b), n=10:
Gjennomsnittsdifferanse mellom to måletidspunkt er vist langs y-aksen (tidspunkt 2 – tidspunkt 1) og gjennomsnitt mellom måletidspunktene langs x-aksen (tidspunkt 1 + tidspunkt 2/2).
LoA (Limits of Agreement) er vist som rød stiplet linje, kalkulert gjennomsnittsdifferanse ±1,96*SD

6. Diskusjon

Denne studien har vist at ventrikkel og subaraknoidal CSF kan automatisk segmenteres til nært nivå av manuell gullstandard, selv med lavt antall treningsdata for KI-algoritmen. I dette kapittelet blir evalueringene, metoden og etikken diskutert, samt en redegjørelse for styrker og svakheter ved studien.

6.1 Validering av KI-algoritmen

Utviklingen av metoden for segmentering av vCSF og sCSF viste at KI-algoritmen som var opptrent med mer enn 25 datasett (versjon 4 trent med 25, og versjon 5 trent med 35) utførte nær gullstandard presisjon. Antall datasett for opptrening i denne studien er veldig lavt, sammenlignet med lignende valideringer av segmenteringsalgoritmer der det kan være over 1000 (33). Graden av manuell korrigering som var nødvendig fra KI-algoritme versjon 4 og 5, påvirket endelig vCSF og sCSF volum i liten grad og var ikke signifikant forskjellig fra totalvolumet. Derfor kan manuell korrigering vurderes å begrenses til det minimale.

Utviklingen til KI-algoritmen fra versjon 1-3 til 4-5 ble tydeligst illustrert ved Dice koeffisienten. ICC og Bland-Altman plott viste høyt samsvar *også* for KI-algoritme versjon 1-3, men hadde usikkerhet knyttet til vCSF ved større konfidensintervall, og til sCSF med stort LoA intervall. Samsvar mellom gjennomsnittsvolumet til KI-algoritmen og gullstandard, var lavere for vCSF og påvirket konfidensintervallet i ICC. Store differanser for sCSF, ga et range fra -13,1 ml til 16,7 ml i LoA, for versjon 1-3, som betyr at det var individuelle store differanser innad i de første tre versjonene av algoritmen.

Årsaken til at piksel-overlapp fikk best frem utviklingen av algoritmen, var at Dice koeffisient er mer sensitiv for feilsegmenteringer, enn volum. Når antall pixler som tilsvarer 1 ml CSF varierer fra 1042 til 4875, kan volumsamsvar tolkes å være mindre nøyaktige. Volumene kan i tillegg gi misvisende samsvar hvis den underestimerer 1000 piksler CSF, men feilregistrerer 1000 piksler hjernematerie i samme datasett og volumet derfor blir likt.

Spredningsplottene som viste piksel-overlapp for hver KI-algoritme versjon, bekreftet at versjonene innad i gruppene 1-3 og 4-5 hadde samme resultater, stor spredning for alle tre første versjoner og markant mindre i de to siste. Dette bekreftet at hoppet i presisjonen forekom fra versjon 3 til versjon 4.

Alderen til barna segmentert i KI-algoritme versjoner 4-5 viste gjennomsnittlig høyere volum, 14,9 ml, enn barna for versjon 1-3, 9,9 ml. Sannsynlig årsak var alderen, med gjennomsnitt på

61 uker for versjon 4-5 og 44 uker for versjon 1-3. Det hadde vært en fordel med mindre spredning i alder mellom versjon 1-3 og 4-5, grunnet potensielle bias med ulik anatomi ved ulike alderstrinn, og at dette påvirket algoritmens presisjon.

6.2 Inter-rater reliabilitet

For gruppen «normal CSF» var det ingen tendenser til systematisk feil mellom raterne, og viste svært høy reliabilitet. Raterne hadde høy, men noe lavere reliabilitet for gruppen «Økt CSF». Det ble observert større konfidensintervall for ICC og høyere gjennomsnittsdifferanser. Bland-Altman plott viste to uteliggere for sCSF, for «økt CSF» med differanser mellom Rater 1 og Rater 2, på ≥ 7 ml, men må ses i forhold til et høyt totalvolum av sCSF på ≥ 180 ml. Med et så høyt totalvolum, er de godt utenfor referanseområdene med gjennomsnitt på 125 ml sCSF hos en voksen (10). I gruppen «økt CSF» var det få observasjoner med høye volum, og er derfor mer usikkerhet knyttet til datasett med større volum.

Rater 2 segmenterte konsekvent større volum av sCSF for gruppen «økt CSF», der alle differansene var over 0. Denne over/under-estimeringen kan tolkes som tendens til systematisk feil, men ble ikke sett i gruppen «normal CSF». Mest sannsynlig har det sammenheng med at raterne hadde mindre erfaring med økt volum. På bakgrunn av lavere samsvar for gruppen «økt CSF» bør raterne øke felles forståelse av gullstandard for hjerner med økt CSF.

Raterne bør diskutere enighet om gullstandard for hjerner med økt CSF, for å hindre systematisk over/under-segmentering, og tilstrebe likt samsvar som for volum av «normal CSF». Forklaringen bak mindre samsvar mellom raterne for «økt CSF» kan i tillegg skyldes at KI-algoritmen underestimerte segmenteringen og flere piksler måtte manuelt korrigeres. Algoritmen blir ikke bedre enn de datasettene den er trent opp med, er den ikke trent opp med anatomi med forstørrede ventrikler, vil den ikke klare å gjenkjenne dette.

6.3 Intra-rater reliabilitet

Intra-rater reliabilitet viste svært høyt samsvar ved de to måletidspunktene, og betyr at tidspunktene for segmenteringene ikke vil utgjøre et stort bias. Datasettene var identiske med «normal CSF» fra inter-rater reliabilitet, og viste høyere gjennomsnittsdifferanse for sCSF, på 0,9 ml mot 0,4 ml ved inter-rater reliabilitet. Bland-Altman plottet viste en differanse som skilte seg ut på over 2 ml, som påvirket gjennomsnittsdifferansen. Dice koeffisienten ble ikke påvirket og var like høy for inter- og intra-rater reliabilitet.

For sCSF ble det på et måletidspunkt konsekvent målt større volum, da alle gjennomsnittsdifferansene var over 0 ml. Volumdifferansene var ellers lave, LoA viste fra -0,4 ml - 2,2 ml, og dermed uten klinisk betydning når totalvolum var tilnærmet 95 ml. Allikevel kan dette potensielt medføre en skjevhet i dataene, hvis målingene vil fortsette å overestimere i større grad.

6.4 Matematisk vurdering av samsvar

Dice koeffisient, som måler den reelle piksel-overlapp, er som tidligere nevnt i metoden, en matematisk evaluering av algoritmen.

Alle datasett var over akseptabel Dice koeffisient grense på 0,7 (43), lavest for KI-algoritme versjon 1-3 på 0,83. I spredningsplottene der KI-algoritme versjonene ble fremstilt separat, var en observasjon fra versjon 1, for sCSF, under grensen, på 0,6. Siden gullstandard var basert på segmentering av KI-algoritme før manuell korrigering, måtte det forventes en høy piksel-overlapp, og grensen på 0,7 blir noe lav for denne studien. Dice-koeffisienten var over 0,96 for alle (unntatt KI-algoritme versjon 1-3) og viste like høy piksel-overlapp som en lignende studie der de har sammenlignet algoritmer basert på KNN og U-Net med manuell segmentering (33).

Antall voksler, som tilsvarer antall piksler segmentert, per ml CSF, varierte i datasettene fra 1042 til 4875. Derfor er piksel-overlapp en mer presis validering av KI-algoritmen. Volumet som genereres, kan være basert på piksler som er «feilsegmentert», det vil si at det er segmentert CSF, der det ikke er CSF.

CSF segmenteringene kan utføres uavhengig av rater eller tidsintervall, vist ved at inter- og intra-rater reliabiliteten viste Dice koeffisient $\geq 0{,}96$ for samtlige observasjoner. Dette betyr både at svært høy pixel-overlapp fører til samsvar for CSF volum, men i tillegg at vi segmenterer *de samme områdene* i datasettene, og derfor er metoden pålitelig.

6.5 Klinisk vurdering av samsvar

Bland-Altman plottene gir ikke svar på akseptabel klinisk differanse, dette må tolkes av forskerne (41). Differansene mellom KI og gullstandard og ved inter- og intra-rater reliabilitet, må ses i sammenheng med totalvolum. vCSF utgjør en mindre prosentandel av total CSF, og det er gir mer mening å snakke om akseptable grenser i prosent.

En differanse på 10 % som vCSF hadde for KI-algoritme versjon 1-3 vs. gullstandard, kan ikke tolkes som klinisk akseptabel i denne sammenhengen fordi det gir moderat påvirkning av

volumet. Differansen på 1 % mellom raterne for vCSF i gruppen «økt CSF» og resterende differanser for vCSF på ≥ 0,7 % kan tolkes som klinisk akseptabelt da det ikke utgjør noen vesentlig forskjell for totalvolumene.
Differansene for sCSF var på det høyeste 3 % (KI-algoritme versjon 1-3), og 2 % («økt CSF») ved inter-rater reliabilitet, og kan tolkes å være i gråsonen for hva som aksepteres. Resterende differanser var ≥ 0,9 % der intra-rater var høyest, inter-rater reliabilitet og KI-algoritme versjon 4-5 var begge ≥ 0,4 %. Høyt samsvar ved Dice koeffisient og ICC, styrker påstanden om at differansene er klinisk akseptable.

For inter- og intra-rater reliabilitet var det ingen klinisk betydning av differansene mellom raterne ved «normal CSF», og basert på dette kan en direkte implikasjon fra studien være at raterne kan utføre korrigeringer uavhengig av hverandre.

En publikasjon (2019) der hensikten var å måle inter- og intra-rater reliabilitet av semi-automatisk segmentering av abdominale CT bilder ved SliceOmatic, var resultatene av ICC ≥ 0,938 for inter-rater og ≥ 0,996 for intra-rater. Konklusjonen var at segmenteringene ikke var påvirket av raterne og viste nær identiske resultater (44). Metoden som ble brukt var ikke basert på KI-algoritme, men kan ses i sammenheng med denne studiens resultater for inter- og intra-rater reliabilitet.

6.6 Gullstandard

I lignende studier for validering av segmenteringsalgoritmer, der det sammenlignes med gullstandard, er denne oftest segmentert utelukkende manuelt (45). Slike studier har som hensikt å validere tilgjengelige segmenteringsalgoritmer eller egenutviklede algoritmer.

Denne studien valgte å definere gullstandard som segmentert først av KI-algoritmen og deretter manuelt korrigert til gullstandard, av flere årsaker. Manuell segmentering av hele hjernen er tidkrevende (22, 46). I vårt materiale brukte vi 8-10 timer per datasett, og ble derfor valgt en løsning som bestod av automatisk segmentering og manuell korrigering.

I tillegg hadde det sammenheng med problemstillingene som skulle undersøkes, som var samsvar mellom KI-algoritmen og det ferdige korrigerte datasettet. Derfor begrenses validiteten av algoritmen seg til denne studien og kan ikke sammenlignes med andre algoritmer som er validert mot rene manuelle segmenteringer. En ren manuell gullstandard ville økt styrken av KI-algoritmens validitet.

6.7 Statistiske analyser

Få observasjoner med n=10 i hver variabel som ble sammenlignet, er langt under anbefalt minimum med n=50 for reliabilitets- og validitets-studier (47). ICC og Bland-Altman plottene må i denne studien tolkes med forsiktighet, grunnet lavt antall datasett. Det ble også sett få observasjoner med store gjennomsnitt (store volum) i flere av Bland-Altman plottene og innebærer enda høyere usikkerhet når det gjelder datasett med større CSF volum.

Et større antall datasett hadde tilført studien større gyldighet. For validitetsevaluering av KI-algoritmen var manuell korrigering allerede utført, men grunnet manglende data av segmenteringer utført av KI-algoritmen, var det nødvendig å redusere utvalget. Siden algoritmen erstattes med ny versjon fortløpende, var det ikke mulig å generere tidligere versjoner. Datasett med rene KI-algoritme segmenteringer som ble benyttet i studien, var blitt lagret på det tidspunktet de ble generert.

De statistiske analysene ble gjennomført som om de var normalfordelte, selv om både histogrammer og QQ-plott ga varierende resultater. Manglende normalfordeling hadde sannsynlig sammenheng med antall datasett per variabel. Gitt at n=100, ville dette mest sannsynlig gitt normalfordeling. Ikke-parametriske data kan ved logaritmiske transformasjoner oppnå normalfordeling, men kan medføre feilaktige konklusjoner da det potensielt medfører nye problemer (36). Ikke-parametriske tester som korrelasjonsanalyse ble ikke vurdert, da det var absolutt samsvar som var målet. Supplementet med Dice koeffisient, som ikke berøres av normalfordelinger, kan argumenteres å minimere usikkerheten til resultatene fra ICC.

6.8 Etiske betraktninger

REK godkjente uthenting av avidentifiserte MR undersøkelser uten samtykke for det overordnede prosjektet, grunnet stor nytteverdi. Materialet i masterstudien har benyttet de samme datasettene, for metodeutvikling og evaluering. For bevaring av autonomien, var informasjon om det overordnede prosjektet publisert på internettsidene for OUS, med mulighet for å reservere seg fra forskningsprosjektet.

Informert samtykke er ikke det som primært gjør medisinsk forskning forsvarlig, men hva deltakerne utsettes for satt opp mot nytte for pasienten, gruppen eller samfunnet for øvrig (48). For det overordnede prosjektet ville innhenting av samtykker vært veldig krevende med tanke på størrelsen av materialet med mål om å danne et referansemateriale. Manglende

samtykker ville utgjort et stort bias i resultatene. På bakgrunn av godkjenningen fra REK ble det vurdert til at masterstudien var en del av dette prosjektet, og alle data ble behandlet med tilsvarende føringer med avidentifiserte datasett.

Metodeutviklingen og evaluering av denne vil vitenskapelig bidra til prosjektet med å samle inn volummålinger til et referansemateriale. I den sammenheng vil det bidra til nytte på individnivå, både pasienter og pårørende. På et samfunnsmessig plan vil et slikt materiale ha stor betydning for ulike fagmiljøer innen barnemedisin og rettsmedisin.

Metoden og resultatene ble fremstilt på mest mulig transparent måte, og resultater presentert i sin helhet, uten påvirkning av egeninteresse.

6.9 Styrker og svakheter ved studien

Masterstudien har gjennomført metodeutvikling med en tosidig evaluering, klinisk med volumsamsvar og matematisk med piksel-overlapp. Evalueringene har vist at metoden gir en valid og reliabel måling av CSF på MR-bilder av barn fra 0 til 2 år.

De fleste automatiske segmenteringsalgoritmer har spesifikke krav til sekvensen, som regel kreves det et 3D sekvens (11, 22). Dette gjør algoritmene mindre egnet for denne typen retrospektive studier med kliniske undersøkelser, fra ulike skannere, og ikke bruker kompatible forskningssekvenser. 3D sekvenser har inntil nylig ikke vært standard i barneprotokoller grunnet lengre opptakstid. De siste årene har det vært en utvikling av opptaksteknikker som korter ned tiden. For denne metodestudien, var vi avhengige av å utvikle en metode for varierte T2 2D sekvenser, da materialet bestod av store variasjoner i kontraster og oppløsning. KI-algoritmen ble trenet opp til å segmentere 2D sekvenser med ulike T2 kontrast-parametre og ulik oppløsning, og å differensiere mellom vCSF og sCSF, den utviklede KI-algoritmen er i den forstand robust.

Studien kan kritiseres for et lite utvalg og det kan ha påvirket resultatene i de statistiske analysene. Utvalget må imidlertid ses i sammenheng med at hvert datasett inneholder 20-40 bilder og manuell korrigering tar tid. Datautvalget påvirket normalfordelingen og gyldigheten til statistiske analysene, e.g. ICC kriteriene var ikke oppfylt og må dermed tolkes deretter. For den matematiske evalueringen kan det argumenteres med at hvert datasett kan multipliseres med antall bilder i datasettet, og dermed gir et høyere kvantum.

En gullstandard basert på utelukkende manuell segmentering hadde økt validiteten til resultatene. Samtidig hadde det også gjort studien mer sammenlignbar med andre studier.

Sett i etterkant kunne denne studien med fordel blitt utført med flere data for færre variabler, eksempelvis kun evaluert KI-algoritmen mot gullstandard, med datasett med både normale og økte CSF volum. Det ville begrenset problemstillingenes omfang, men ville medført høyere grad av validitet. Med bakgrunn i den høye presisjonen KI-algoritmen viste fra og med versjon 4, kan manuell korrigering vurderes og begrenses til det minimale, og det innebærer at inter- og intra-rater reliabiliteten fremover ikke vil påvirke segmenteringen i like stor grad. En fordel med automatisk segmentering, utenom tidsbesparende, er fraværet av påvirkning fra ratere (9).

Datasettene som ble inkludert var fra barn som hadde foretatt MR-undersøkelse av ulike kliniske årsaker. Med inklusjonskriteriene om negativt radiologisvar mener prosjektet at de kan bli sett som en normalpopulasjon. Dette kan medføre et mulig bias siden de var henvist til MR for en grunn. Metoden ble utviklet på et materiale med potensielle variasjoner fra en normalpopulasjon.

7. Videre forskning og fremtidsperspektiver

I lys av de metodiske begrensningene med gullstandard basert på KI-algoritme segmentering, hadde det vært interessant og foretatt tilsvarende analyser med utelukkende manuelt segmentert gullstandard. 10 datasett hadde vært tilstrekkelig for å se om dette ga like høyt validitet som evalueringene i masterstudien.

Gullstandarder som er basert på faglig ekspertise, er utsatt for subjektivitet, og derfor er det laget en algoritme for å lage en mer objektiv gullstandard, STAPLE (49). STAPLE-algoritmen bearbeider flere enn to datasett, eksempelvis to manuelt segmentert og en algoritmesegmentert. Deretter kalkuleres et probabilistisk estimat av «sann» segmentering. En videreutvikling av metoden for denne studien med STAPLE, ville bidratt med nyutvikling innen feltet og en mer validert gullstandard.

En annet interessant retning ville vært å teste KI-algoritmen med andre validerte algoritmer. 3D sekvenser *kan* segmenteres av KI-algoritmen, da de genererer reformater i aksialt plan som KI-algoritmen tolker som opptrente 2D sekvenser. Evalueringer med volumsamsvar og Dice koeffisient kunne blitt brukt for sammenligning med en annet automatisk segmenteringsprogram.

8. Konklusjon

I denne studien har vi utviklet og evaluert en metode basert på kunstig intelligens, for segmentering av ventrikkel CSF og subaraknoidal CSF. Utviklingen av metoden har vært en trinnvis prosess fra manuell til automatisk segmentering, og gjennom KI-algoritmens fem versjoner. Evalueringen viste en tydelig positiv utvikling fra de tre første, til de to siste KI-algoritme versjonene. De to siste versjonene viste svært høye resultater for validitet, med økt presisjon nært gullstandardnivå, og kan føre til at den manuelle korrigeringen reduseres til det minimale. Inter-rater reliabiliteten var svært høy for gruppen med normale CSF volum, og moderat/høy ved økt CSF volum. Intra-rater reliabiliteten var svært høy. Funnene av konsekvent mer segmentering ved inter-rater reliabilitet ved «økt CSF» og ved intra-rater reliabilitet, bør diskuteres videre av raterne for å unngå systematiske feil.

Den validerte KI metoden kan bli brukt for segmentering av et større materiale, innsamling av et referansemateriale, for CSF volum gjennom de to første leveårene.

Referanser

1. Lynøe N, Elinder G, Hallberg B, Rosén M, Sundgren P, Eriksson A. Insufficient evidence for 'shaken baby syndrome' - a systematic review. Acta Paediatr. 2017;106(7):1021-7.
2. Saunders D, Raissaki M, Servaes S, Adamsbaum C, Choudhary AK, Moreno JA, et al. Throwing the baby out with the bath water — response to the Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU) report on traumatic shaking. Pediatric Radiology. 2017;47(11):1386-9.
3. Stray-Pedersen A, Møller C, de Lange C, Due-Tønnessen BJ, Grøgaard JB, Haugen OH, et al. The doctors' role in cases of suspected child abuse. Tidsskr Nor Laegeforen. 2019;138(2).
4. Thiblin I, Andersson J, Wester K, Wikström J, Högberg G, Högberg U. Medical findings and symptoms in infants exposed to witnessed or admitted abusive shaking: A nationwide registry study. PLOS ONE. 2020;15(10):e0240182.
5. Zahl SM, Egge A, Helseth E, Wester K. Benign external hydrocephalus: a review, with emphasis on management. Neurosurg Rev. 2011;34(4):417-32.
6. Marino MA, Morabito R, Vinci S, Germanò A, Briguglio M, Alafaci C, et al. Benign external hydrocephalus in infants. A single centre experience and literature review. Neuroradiol J. 2014;27(2):245-50.
7. Vinchon M, Delestret I, DeFoort-Dhellemmes S, Desurmont M, Noulé N. Subdural hematoma in infants: can it occur spontaneously? Data from a prospective series and critical review of the literature. Child's Nervous System. 2010;26(9):1195-205.
8. Kahle KT, Kulkarni AV, Limbrick DD, Jr., Warf BC. Hydrocephalus in children. Lancet. 2016;387(10020):788-99.
9. Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications. Comput Math Methods Med. 2015;2015:450341.
10. Sakka L, Coll G, Chazal J. Anatomy and physiology of cerebrospinal fluid. Eur Ann Otorhinolaryngol Head Neck Dis. 2011;128(6):309-16.
11. Serai SD, Dudley J, Leach JL. Comparison of whole brain segmentation and volume estimation in children and young adults using SPM and SyMRI. Clin Imaging. 2019;57:77-82.
12. Sharma N, Aggarwal LM. Automated medical image segmentation techniques. J Med Phys. 2010;35(1):3-14.
13. Gilmore JH, Knickmeyer RC, Gao W. Imaging structural and functional brain development in early childhood. Nature Reviews Neuroscience. 2018;19(3):123-37.
14. Tumani H, Huss A, Bachhuber F. Chapter 2 - The cerebrospinal fluid and barriers – anatomic and physiologic considerations. In: Deisenhammer F, Teunissen CE, Tumani H, editors. Handbook of Clinical Neurology. 146: Elsevier; 2018. p. 21-32.
15. Greitz D. Radiological assessment of hydrocephalus: new theories and implications for therapy. Neurosurg Rev. 2004;27(3):145-65; discussion 66-7.
16. Iliff JJ, Wang M, Liao Y, Plogg BA, Peng W, Gundersen GA, et al. A Paravascular Pathway Facilitates CSF Flow Through the Brain Parenchyma and the Clearance of Interstitial Solutes, Including Amyloid β. Science Translational Medicine. 2012;4(147):147ra11-ra11.
17. Ringstad G, Vatnehol SAS, Eide PK. Glymphatic MRI in idiopathic normal pressure hydrocephalus. Brain. 2017;140(10):2691-705.
18. Abildgaard A. MR for radiografer og radiologer : fysikk og fysiologi. Oslo: Universitetsforl.; 2016.
19. Westbrook C, Roth CK, Talbot J. MRI in Practice: Wiley; 2011.
20. Zou KH, Wells WM, 3rd, Kikinis R, Warfield SK. Three validation metrics for automated probabilistic image segmentation of brain tumours. Stat Med. 2004;23(8):1259-82.
21. Lucena O, Souza R, Rittner L, Frayne R, Lotufo R. Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks. Artif Intell Med. 2019;98:48-58.

22. Grøvik E, Yi D, Iv M, Tong E, Nilsen LB, Latysheva A, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study. npj Digital Medicine. 2021;4(1):33.
23. Weisenfeld NI, Warfield SK. Automatic segmentation of newborn brain MRI. NeuroImage. 2009;47(2):564-72.
24. Hyde DE, Duffy FH, Warfield SK. Anisotropic partial volume CSF modeling for EEG source localization. NeuroImage. 2012;62(3):2161-70.
25. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69s:S36-s40.
26. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. Journal of Medical Imaging and Radiation Sciences. 2019;50(4):477-87.
27. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep Learning: A Primer for Radiologists. RadioGraphics. 2017;37(7):2113-31.
28. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60-88.
29. Gotra A, Sivakumaran L, Chartrand G, Vu KN, Vandenbroucke-Menu F, Kauffmann C, et al. Liver segmentation: indications, techniques and future directions. Insights Imaging. 2017;8(4):377-92.
30. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-29.
31. Kido S, Hirano Y, Mabu S. Deep Learning for Pulmonary Image Analysis: Classification, Detection, and Segmentation. Adv Exp Med Biol. 2020;1213:47-58.
32. 1. Tomovision [Internett] Magog Ch--Tfhwtcpsh.
33. Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, et al. Automated Abdominal Segmentation of CT Scans for Body Composition Analysis Using Deep Learning. Radiology. 2019;290(3):669-79.
34. Fischl B. FreeSurfer. Neuroimage. 2012;62(2):774-81.
35. Krithikadatta J. Normal distribution. J Conserv Dent. 2014;17(1):96-7.
36. Juneja P, Evans PM, Harris EJ. The validation index: a new metric for validation of segmentation algorithms using two or more expert outlines with application to radiotherapy planning. IEEE Trans Med Imaging. 2013;32(8):1481-9.
37. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.
38. Plichta SB, Kelvin EA, Munro BH. Munro's statistical methods for health care research. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2012.
39. Carter RE, Lubinsky J, Domholdt E, Domholdt E. Rehabilitation research : principles and applications. 2011.
40. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-51.
41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
42. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil. 1998;12(3):187-99.
43. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Transactions on Medical Imaging. 1994;13(4):716-24.
44. Kjønigsen LJ, Harneshaug M, Fløtten AM, Karterud LK, Petterson K, Skjolde G, et al. Reproducibility of semiautomated body composition segmentation of abdominal computed tomography: a multiobserver study. Eur Radiol Exp. 2019;3(1):42.
45. Grimm O, Pohlack S, Cacciaglia R, Winkelmann T, Plichta MM, Demirakca T, et al. Amygdalar and hippocampal volume: A comparison between manual segmentation, Freesurfer and VBM. J Neurosci Methods. 2015;253:254-61.

46. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. Semin Radiat Oncol. 2019;29(3):185-97.
47. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide: Cambridge University Press; 2011.
48. magelssen.
49. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23(7):903-21.

Vedlegg 1

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK sør-øst	Tove Irene Klokk	22845522	08.02.2019	2018/2510/REK sør-øst
				A
			Deres dato:	Deres referanse:
			11.12.2018	

Vår referanse må oppgis ved alle henvendelser

Heidi Beate Eggesbø
Forsknings- og utviklingsavdelingen (FoU)

2018/2510 Cerebrospinalvæskevolum gjennom de to første leveårene

Forskningsansvarlig: Oslo universitetssykehus HF **Prosjektleder:** Heidi Beate Eggesbø

Vi viser til søknad om forhåndsgodkjenning av ovennevnte forskningsprosjekt. Søknaden ble behandlet av Regional komité for medisinsk og helsefaglig forskningsetikk (REK sør-øst) i møtet 17.01.2019. Vurderingen er gjort med hjemmel i helseforskningsloven (hforsknl) § 10.

Prosjektbeskrivelse (revidert av REK)

Hjernevæsken er væsken som ligger mellom skallen og hjernen samt i hulrom inne i hjernen. Så vidt prosjektleder vet, finnes det ingen studier som har kartlagt normalverdier og endringer i hjernevæskevolum gjennom de to første leveårene på levende barn. Et slikt normalmateriale vil være et nyttig hjelpemiddel i vanskelige eller utfordrende diagnosesituasjoner. Sykdommer og tilstander som hjernehinnebetennelse, hjernebetennelse, hodeskade (ulykke eller påført) og blødninger kan påvirkemengden av hjernevæske.

Hensikten med denne studien er derfor å utføre volummålinger av hjernevæske hos barn som allerede har vært henvist til og gjennomført MR av hodet av ulike årsaker. Basert på nye analyser av MR bildene, ønsker forskerne å kartlegge:

1) Normalverdier for hjernevæskevolum for barn 0-2 år, ved å måle indre og ytre hjernevæske på bilder frautførte MR undersøkelser.
2) Hjernevæskevolum for barn 0-2 år hos barn med ulike diagnoser, med referanse i normalmaterialet frapunkt 1.
3) Forekomsten av godartet ytre vannhode hos norske spedbarn

I tillegg til nye analyser av MR bilder, skal det innhentes helseopplysninger om barna fra radiologisk journalsystem (RIS/PACS), det medisinske journalsystemet ved Oslo Universitetssykehus (DIPS) og helsestasjonsrapporter.

Det skal hentes opplysninger fra omtrent 3000 barn under 2 år, som allerede har tatt MR caput, i tiden fra 2011 til i dag. Disse skal grupperes etter henvisningsårsak og diagnose. Det bes om fritak fra samtykke for innhenting av disse opplysningene.

Vurdering

Komiteen mener dette er et nyttig og godt begrunnet prosjekt, med potensiell stor samfunnsnytte. Det å

Besøksadresse: Gullhaugveien 1-3, 0484 Oslo
Telefon: 22845511
E-post: post@helseforskning.etikkom.no
Web: http://helseforskning.etikkom.no/

etablere et normalvolum for cerebrospinalvæsken hos barn, vil ha stor nytteverdi ved diagnostiseringen av sykdommer og skader som påvirker hjernen.

Forskerne skal innhente opplysninger fra omtrent 3000 barn som fra 2011 fremt til i dag har gjennomført MR undersøkelse av hodet ved Oslo Universitetssykehus (OUS). Opplysningene som skal hentes ut er:

Fra helsestasjonsrapporter:
- Hodeomkrets
- Fødselsvekt
- Psykomotorisk og kognitiv utvikling

Fra pasientjournal (RIS/PACS, DIPS og helsestasjonsrapporter):
- Henvisningsårsak
- Kjønn og alder
- Endelig diagnose

Det skal ikke innhentes ny informasjon eller foretas nye undersøkelser av deltakerne. Forskerne skal derimot hente ut MR bilder som er tatt av pasientene, og analysere disse for hjernevæskevolum, Evans index, hodeomkrets, funn i hjernen som kan passe med ischemisk hjerneskade, og andre relevant MR-funn i hjernen.

Det søkes om fritak fra samtykke for innhenting av de nevnte opplysninger og bruk av MR - bilder. Det argumenteres med at det ikke skal samles inn nye opplysninger og at materialet vil bli forløpende avidentifisert av prosjektleder. Det vektlegges at det kan det være utfordrende å innhente samtykke hos spesielle grupper (omsorgssvikt/mishandling), og at et eventuelt frafall kan gi bias i resultatene og dermed reduserer kvaliteten på arbeidet.

REK kan bestemme at helseopplysninger innsamlet i helse- og omsorgstjenesten kan utleveres til bruk i forskning, og at det kan skje uten hinder av taushetsplikt, jf. helseforskningsloven § 35. Dette kan bare skje dersom slik forskning er av vesentlig interesse for samfunnet og hensynet til deltakernes velferd og integritet er ivaretatt. Den regionale komiteen for medisinsk og helsefaglig forskningsetikk kan sette vilkår for bruken, blant annet for å verne de registrertes grunnleggende rettigheter og interesser.

Bestemmelsen må tolkes tilsvarende helseforskningslovens §§ 15 annet ledd og 28 første ledd. I praksis betyr dette at det også skal være vanskelig å innhente nytt samtykke.

Komiteen er enig i at prosjektet har vesentlig samfunnsnytte, da et referansevolum for cerebrospinalvæsken hos barn vil ha stor betydning for ulike fagmiljøer som spesialister i barnemedisin og rettsmedisin. Deltakernes velferd og integritet er ivaretatt gjennom avidentifisering av opplysningene. Å skulle innhente samtykke fra 3000 pasienter anses som overkommelig, men medfører vesentlig merarbeid. Komiteen godkjenner derfor at prosjektet gjennomføres uten at det innhentes samtykke.

Komiteen mener likevel at deltakerne har rett til å få vite at deres helseopplysninger brukes til forskning, og ha mulighet til å reservere seg mot dette. Dette gjelder spesielt siden dette dreier seg om en sårbar gruppe pasienter, barn 0-2 år. Dette kan gjøres ved at det informeres om prosjektet på OUS/avdelingens hjemmesider, hvor det også opplyses om retten til å reservere seg og hvordan man eventuelt gjør dette.

Komiteen godkjenner derfor med hjemmel i helseforskningsloven § 35 at data som beskrevet i søknad og protokoll blir benyttet i prosjektet, på vilkår om at informasjon om prosjektet gjøres tilgjengelig og det opplyses om retten til å reservere seg fra at deres barns helseopplysninger blir benyttet i forskning.

Prosjektleder er ført opp som kontaktperson ved forskningsansvarlig institusjon. Prosjektleder kan ikke være samme person som institusjonens kontaktperson. Komiteen gjør oppmerksom på at kontaktperson ved forskningsansvarlig institusjon skal være institusjonens øverste leder, eller den som øverste leder har delegert oppgaven til. Det bes om at det sendes inn navn, stilling og e-post adresse til ny forskningsansvarlig person.

Etter en helhetlig vurdering, setter komiteen følgende vilkår for godkjenning av prosjektet:

- Det skal informeres om prosjektet på OUS sine hjemmesider, inkludert informasjon om retten til åreservere seg fra bruk av deres barns helseopplysninger i forskning.
- Det må sendes inn navn, stilling og e-post til ny forskningsansvarlig person.

Vedtak

REK har gjort en helhetlig forskningsetisk vurdering av alle prosjektets sider. Prosjektet godkjennes med hjemmel i helseforskningsloven § 10, under forutsetning av at ovennevnte vilkår er oppfylt.

Vi gjør samtidig oppmerksom på at etter ny personopplysningslov må det også foreligge et behandlingsgrunnlag etter personvernforordningen. Det må forankres i egen institusjon.

I tillegg til vilkår som fremgår av dette vedtaket, er godkjenningen gitt under forutsetning av at prosjektet gjennomføres slik det er beskrevet i søknad og protokoll, og de bestemmelser som følger av helseforskningsloven med forskrifter.

Med hjemmel i helseforskningsloven § 35 gir komiteen fritak fra samtykkekravet, herunder dispensasjon fra taushetsplikten, for bruk av helseopplysninger innsamlet i helsetjenesten til forskningsformål slik det er beskrevet i søknaden.

Komiteens avgjørelse var enstemmig. Godkjenningen gjelder til 31.12.2031.

Av dokumentasjonshensyn skal opplysningene oppbevares i 5 år etter prosjektslutt. Opplysningene skal oppbevares avidentifisert, dvs. atskilt i en nøkkel- og en datafil. Opplysningene skal deretter slettes eller anonymiseres.

Prosjektet skal sende sluttmelding på eget skjema, jf. helseforskningsloven § 12, senest et halvt år etter prosjektslutt.

Dersom det skal gjøres endringer i prosjektet i forhold til de opplysninger som er gitt i søknaden, må prosjektleder sende endringsmelding til REK, jf. helseforskningsloven § 11.

Klageadgang
Komiteens vedtak kan påklages til Den nasjonale forskningsetiske komité for medisin og helsefag, jf. helseforskningsloven § 10 tredje ledd og forvaltningsloven § 28. En eventuell klage sendes til REK sør-øst A. Klagefristen er tre uker fra mottak av dette brevet, jf. forvaltningsloven § 29.

Vi ber om at alle henvendelser sendes inn på korrekt skjema via vår portal: https://helseforskning.etikkom.no. Dersom det ikke finnes passende skjema kan henvendelsen rettes på epost til: post@helseforskning.etikkom.no.

Med vennlig hilsen

Knut Engedal
Professor dr. med.
Leder

Tove Irene Klokk
Rådgiver

Kopi til: h.b.eggesbo@medisin.uio.no; Oslo universitetssykehus HF ved øverste administrative ledelse: oushfdlgodkjenning@ous-hf.no

Vedlegg 2

CSF– *Fra uthenting av data til ferdig korrigerte datasett*

Koder genereres av prosjektleder i forkant.

1. Uthenting av bilder (prosjektleder)

- Fødselsnummer søkes opp i RIS/PACS.
  Hvis pasient har flere MR undersøkelser, skal den første utførte velges.
  Noteres i Filemaker at flere datasett er tilgjengelige.
- Følgende noteres i Filemaker:
- Henvisningsårsak
- Radiologisk svar (positiv/negativ – obs lett markert subaracnoidalt rom = «diagnose»)
- Bildekvalitet (bevegelse, oppløsning, hele hjernen er med)

2. SliceOmatic maskin (overføre bilder)

- Bildesett overføres fra IronKey til mappen **Rådata** (**LACIE: (E:) - Hjerne_hjerte - MR caput**)
- Kopier datasett fra **rådata** over til mappen **1_AI input tomme TAGfiles**
  (**MR caput – AI_segmentering – AI prosess CSF**).
- Snitt under Foramen Magnum må slettes (gjør dette før segmenteringsprosess).

3. Lag tomme tagfiles

- Åpne **SliceOmatic (Alberta protocol – kun denne brukes),**
  feilmeldinger dukker opp automtaisk – kryss ut.
- Åpne datasett fra **1_AI input tomme TAGfiles** i SliceOmatic via File – Dicom browser
  eller drag & drop fra 1_AI input tomme TAGfiles.
- Sjekk at snitt under Foramen Magnum er slettet, hvis ikke, gjør dette nå.
- Gå via Modes (høyre øvre hjørne)- **step 3: Segmentation**.
- Trykk **Grow 3D** i verktøyvindu (---Region Growing---) marker tynneste pensel –
  klikk ON på lower og upper limit. Dra upper limit terskelboks max mot høyre.
- Klikk i bildet, den markeres nå i lilla.
- **Kontroller** at tomme tagfiles (doble bildesett) ligger i riktig mappe
  (1 AI input tomme TAGfiles).
- **Kontroller** at ID kode 4 siffer stemmer med mappe navn: Via File – config – default overlays –
  kryss av Name, ID kode kommer nå opp i bildene.

4. AI segmentering

- Åpne mappe på skrivebord **MR_CSF_AI_segm_v0.0**.
  Input/output/hjernevolum må være tomme.
- Tomme tagfiles fra **1_AI input Tomme TAGfiles** kopieres til **input**.

- Filter kan korrigeres i script på forhånd (0.1-0.9) – via run – cutoff = 0.8. Endre kun tall og trykk save. Høyere tall = mindre segmentering. Husk, mer jobb å viske bort, enn å legge til.
- Trykk **StartAIseg** (algoritme).
  Etter 15 sek ligger det outputprediksjoner i output (CSF) og outputhodevol.
- **Klipp** output CSF over til 2_AI output prediksjoner (lag først mappe til denne koden) og **kopier mappen** (eks 1039) til 3_AI korreksjon. Dermed beholder vi prediksjonene og korrigerer **KUN** fra 3_AI korreksjon.
- Gjenta punkt over med hode volum. Nå skal det ligge duplikater i både 2_AI output og 3_AI korreksjon for CSF og hodevolum.

5. Korrigering CSF

- Åpne SliceOmatic (sOm) – dra bildesett fra **3_AI CSF korreksjon** over i sOm. Trykk space for å skifte mellom et snitt og multiple snitt. Modes – segmentation.
  Marker lilla boks i verktøyvindu (CSF). Marker ON på upper og lower limit.
- Dra upper limit helt til høyre (max GLI, grey level index, verdi). Lower limit justeres kontinuerlig etter hvilke gråtoner som CSF har.
- 2 valg: **Paint** eller **Grow 2D**.
  Paint: bruk ønsket penseltykkelse, ofte lettere med tykk. Når du fører pensel over anatomi ser du hvilket område den inkluderer (ses som rosa).
  Grow 2D: marker tynneste pensel. Klikk i området med CSF og den fyller ut alle gråtonene som terkelverdien er satt til.
- **Viske ut feilsegmenteringer**: marker **none** i verktøyvindu. Nylige feil kan viskes ut ved å høyreklikke på mus.
- **Obs**. mye flow artefakter og pve ved skallebasis. Kontroller snitt over/under ved usikkerhet omkring pve. Store arterier skal unngås å fargelegge. Sjekk tidligere segmenteringer ved behov.
- Programmet autosaver etter noen minutter, men kan gjerne trykke save før du lukker.
- Korreksjonsbildene skal ses over av annen prosjektmedarbeider før vi regner volum.
  Endre filnavn i prediksjon med forbokstaven din. 2. reader setter egen forbokstav etter korreksjonslesing. Eks. CSF 1039_E_B (korrigert av Ellen og lest over av Bianca).
- Ferdig korrigerte datasett (sjekket av to personer) lagres i mappe 4_Brain_CSF_finished.

6. Volum

- Kun valgt serie må være oppe i sOm. Modes - Segmentation må være merket, trykk så step 4: save the results. I boksen som kommer opp trykk configuration:
  Volum cm^3 og **volume voxel** må være huket av (dvs. både ml/cm^3 og antall voxler).
- Excelfil lagres som CSF_kode i mappe **Results** som ligger i mappen **4_Brain_CSF_finished** og som Brainvol_kode i **4_Brain_Volume_finished**.
- Volumdata skal føres inn i skjema volum excelfil under mappe EXCEL_volum og tekniske data under Tekniske_data_sekvenser.
- Når vi har implementert FileMaker kan dette føres direkte dit.