Augmented Reality for EVD Placement:
Evaluating the Accuracy and Clinical Feasibility of Anatomical Landmark Registration.

Merel Goossens

AI-based image generation tools were used to create the illustration for the cover

Augmented Reality for EVD Placement:
Evaluating the Accuracy and Clinical Feasibility of Anatomical Landmark Registration.

Merel Charissa Goossens

Student number: 4856902

January 7, 2026

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)
Dept. of Neurosurgery, UMC Utrecht
May 19, 2025 – January 22, 2026

Supervisor(s):
Dr. T.P.C. van Doormaal, UMC Utrecht
Dr. N.E.C. van Klink, UMC Utrecht
Dr. R.P.J. van den Ende, LUMC

Thesis committee members:
Dr. T. van Walsum, Erasmus MC (chair)
Dr. T.P.C. van Doormaal, UMC Utrecht
Dr. N.E.C. van Klink, UMC Utrecht
Dr. R.P.J. van den Ende, LUMC

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Preface & Acknowledgements

With this thesis, my seven years of studying come to an end. Seven years ago, I was very much in doubt about which path to choose. Should I pursue a technically oriented study, as I had always excelled in technical subjects in high school, or a health-oriented study, since the human body had always intrigued me? It will come as no surprise that I applied for both Medicine and Technical Medicine, and even until the day I received the results, I had not made a clear preference. Secretly, I hoped the choice would be made for me, and that is exactly what happened.

I first completed a three-year bachelor's degree in Medicine, which I enjoyed very much. However, something kept nagging at me. I felt that I was not using my brain in all the ways that I could. I missed approaching problems numerically and through multiple logical steps, rather than mainly memorising information. After a year of doubt, I decided to deviate from the expected path of becoming a physician. Following a pre-master's and a master's program in Technical Medicine, I can now confidently say that this was the right choice.

I enjoyed the diversity of the master's internships to the fullest. I was still able to experience medicine-related activities, while also diving into the development of technical solutions, ranging from working with neural networks to sustainability-related challenges.

From a young age, the world of surgery has inspired me. I remember watching medical TV series and being fascinated by the advanced 3D models and imaging techniques used to plan surgeries in what seemed like futuristic ways. Years later, this future feels much closer. Through my internships, I was introduced to the field of image-guided surgery, and I am grateful for the opportunity to conclude my studies with this graduation internship, in which I created 3D patient models using augmented reality.

I would like to thank Tristan for his willingness to take on the supervision of this project on short notice and for his guidance throughout the eight months of this work. But also Nicole and Tessa for their support and time, and Roy for his valuable fresh perspective. Furthermore, I am grateful for the opportunity to have gained experience at the Amsterdam UMC while contributing to their ongoing research, and I would also like to thank the team at Augmedit. Working closely with a company, which is not a given for our graduation projects, provided quick support for solving technical problems and gave me valuable insight into the clinical implementation of new software. Last but not least, I would like to thank my family and friends for their unwavering support throughout these years.

This thesis marks the end of one journey, but also the beginning of a new one, and I look forward to what lies ahead.

Merel Goossens,
Diemen, January 2026

Abstract

Objective: Accurate placement of external ventricular drains (EVDs) is achieved in only approximately 67–74% of cases using the conventional freehand technique. Augmented reality (AR) offers the potential to improve this by providing real-time, patient-specific anatomical guidance. This thesis evaluates whether CT-based anatomical landmark registration using the Lumi AR workflow is sufficiently accurate, robust, and feasible to support and eventually improve EVD placement. This also includes exploring the clinical acceptability of AI-generated landmarks to streamline the workflow.

Methods: Two studies were performed. First, four clinicians assessed the accuracy of AI-generated anatomical landmarks on CT-derived 3D models, with adjustment rates and interobserver agreement quantified. Second, a prospective pilot study in the operating room (OR) was conducted using the Lumi AR workflow on the HoloLens 2 to perform point-based registration with manually annotated landmarks. The primary outcome was target registration error (TRE); secondary outcomes included fiducial registration error (FRE), visual accuracy ratings, registration time, system robustness and workflow feasibility.

Results: AI-generated landmarks required adjustment in 22.9% of cases (95% CI, 19.1–27.1%), with high median partial interobserver agreement (100.0%, IQR 25.0%) but only moderate mean unanimous agreement (61.0%, 95% CI 51.4–69.7%; Fleiss' kappa = 0.42). In the OR pilot (n=11), the mean TRE at the nasion was 4.9 mm (SD, 2.1 mm). For fiducial validation points, mean TREs were 7.4 mm (SD, 1.7 mm) and 4.9 mm (SD, 1.9 mm). The mean FRE was non-inferior to that reported in a previous phantom study, visual accuracy ratings indicated good perceived alignment, and registration was completed in five minutes on average. Workflow interruptions were primarily due to hardware instability, including three critical failures.

Discussion & Conclusion: AI-generated anatomical landmarks are not yet sufficiently reliable for clinical use in high-stakes scenarios such as EVD placement. In contrast, point-based registration with manually annotated landmarks, using the Lumi AR workflow, proved clinically feasible and achieved an accuracy that is likely acceptable for EVD guidance. However, system robustness remains a key limitation, with AR hardware instability representing the primary obstacle to clinical implementation. Additional limitations include the small pilot sample size, which restricts generalisability, and the variability of soft-tissue surface landmarks. While further advances in AR hardware and validation in larger cohorts are required, these findings indicate that CT-based anatomical landmark registration using AR shows clear potential for guiding future EVD placements.

Table of contents

List of Figures and Tables

List of Abbreviations

AI	Artificial Intelligence
AIOS	Arts In Opleiding tot Specialist (resident in specialist training)
ANIOS	Arts Niet In Opleiding tot Specialist (physician not in specialist training)
AR	Augmented Reality
CE	Conformité Européenne
CI	Confidence Interval
CT	Computed Tomography
CTncSF	CT non-contrast Cranial Segmentation Function
EM	Electromagnetic
EVD	External Ventricular Drain
FLE	Fiducial Localisation Error
FOV	Field Of View
FRE	Fiducial Registration Error
HL2	HoloLens 2
ICU	Intensive Care Unit
IQR	Interquartile Range
METC	Medical Ethics Review Committee
MRI	Magnetic Resonance Imaging
OR	Operating Room
PACS	Picture Archiving and Communication System
SD	Standard Deviation
TRE	Target Registration Error
UMC	University Medical Centre
WMO	Wet medisch-wetenschappelijk onderzoek met mensen

I. Introduction

Neuronavigation is a cornerstone of modern neurosurgery because it helps surgeons determine their exact position within the brain during surgery, much like using GPS or a detailed map to reach a destination. It enhances surgical precision, safety, and patient outcomes by providing surgeons with detailed anatomical views from preoperative imaging, such as computed tomography (CT) or magnetic resonance imaging (MRI) (1-3). While most standard navigation approaches rely on dedicated surgical suites and rigid registration protocols, there is a need for guidance systems that are accurate yet adaptable to urgent and also non-sterile environments. This challenge will be explored in more depth throughout this introduction.

1.1 LIMITATIONS OF FREEHAND EVD PLACEMENT

External ventricular drain (EVD) placement is an example of a routine neurosurgical procedure that could greatly benefit from improved guidance. It is one of the most frequently performed and often lifesaving procedures in neurosurgery, with a prevalence of more than 20,000 in the United States annually (4, 5). EVDs are routinely used to monitor and manage elevated intracranial pressure, particularly in patients with primary hydrocephalus or hydrocephalus secondary to conditions such as subarachnoid haemorrhage, traumatic brain injury, intracerebral or intraventricular haemorrhage or brain tumours (6, 7). The procedure is commonly performed by residents under urgent circumstances, most often in the operating room (OR) or intensive care unit (ICU) and usually relies on freehand techniques guided by anatomical landmarks (7-9). The most widely used approach is via the frontal Kocher's point, which is located approximately 11 cm posterior to the nasion and 3–4 cm lateral to the midline (7).

Despite its widespread use, this technique carries a significant risk of suboptimal placement. Misplacement can lead to serious iatrogenic complications, including haemorrhage, inadequate drainage and nosocomial infections, often requiring revision procedures and consequently increasing patient morbidity and healthcare costs (10). To assess and standardise drain positioning, the Kakarla grading system is frequently used in the literature. This system categorises placement accuracy into three grades: Grade I indicates optimal placement entirely within the ipsilateral frontal horn or tip of the third ventricle, Grade II reflects functional but suboptimal positioning in non-eloquent tissue, and Grade III represents inaccurate placement in eloquent tissue (6). Using freehand techniques, optimal placement (Kakarla Grade I) is achieved in only about 67-74% of cases (11-13). Therefore, there is a clear need for guidance solutions that improve the safety and accuracy of EVD placement.

1.2 REAL-TIME IMAGE GUIDANCE

The substantial risk of misplacement associated with the freehand approach has driven the development and adoption of various real-time guidance methods for ventricular puncture (10, 13). A recent scoping review of 17 studies, including 724 guided procedures, reported consistently favourable outcomes for guided EVD placement. Overall, guided techniques achieved an optimal placement rate of 93.0%, with only 1.1% resulting in the most severe suboptimal placements (Kakarla Grade III). The review identified three main categories of guidance that demonstrated meaningful clinical outcomes (14):

- **Stereotactic Neuronavigation**, mainly encompassing electromagnetic (EM) tracking, is the approach that provided the highest level of accuracy. EM guidance reported a Kakarla Grade I of 93.9% and the lowest rate of dangerous non-functional placements (Kakarla Grade III: 0.9%). However, stereotactic systems require bulky external equipment and time-consuming setup and registration, creating logistical challenges in emergency or non-sterile settings.
- **Ultrasound Guidance** is a portable option offering real-time visualisation through burr-hole or phased-array probes, which achieved Kakarla Grade I rates between 88.5% and 100.0%. Its performance, however, is operator-dependent, relying on the clinician's skill in interpreting a 2D image to align the 3D trajectory.
- **Mechanical Aiming Guides** are simple, cost-effective physical devices that use fixed or adjustable trajectories to guide the drain into the ventricle. These tools showed Kakarla Grade I placement rates ranging from 84.5% to 100.0%. Accuracy was significantly higher (93.0%–100.0%) when preoperative imaging was incorporated into trajectory planning, underscoring the benefit of individualised guidance over fixed-angle approaches.

In conclusion, the overall trend from this review, despite limitations such as heterogeneous data and retrospective data conversion, suggests that guidance methods offer clinically meaningful benefits in achieving optimal drain positioning and may also help reduce the number of insertion attempts. While stereotactic systems showed the highest accuracy, their need for bulky equipment and dedicated setup creates logistical challenges in urgent settings. In contrast, mechanical guides and ultrasound offer simpler, more rapidly deployable options, but at the potential expense of accuracy (14). Therefore, when considering procedures such as EVD placement, there is a need for a guidance solution that combines adequate accuracy with rapid deployability, ergonomic use, and adaptability to diverse clinical environments.

1.3 AUGMENTED REALITY

Recently, augmented reality (AR) has emerged as a promising alternative in surgical navigation to bridge the gap between the need for accurate guidance and the constraints of time-sensitive clinical workflows. Many AR-based navigation systems use head-mounted devices equipped with RGB and depth-sensing cameras, allowing surgeons to view

stereoscopically overlaid virtual anatomical structures aligned with the patient's anatomy in real time (15, 16). This eliminates the need for bulky external equipment and enables surgeons to gain a better spatial and anatomical understanding of the surgical field. AR systems thus also provide ergonomic benefits by reducing the need to shift focus to external displays and minimising physical strain (17, 18).

1.3.1 The Lumi software

Lumi (Augmedit, Naarden, The Netherlands) is a cloud-based AR tool developed in Unity (Unity Technologies, San Francisco, CA, USA) and designed for the Microsoft HoloLens 2 (HL2; Microsoft Corporation, Redmond, WA, USA). It enables clinicians to transform medical imaging data into patient-specific 3D models that support surgical planning and, in the future, intraoperative guidance. Preparation for a procedure begins on the dedicated web application, where 3D patient models (holograms) are generated from CT or MRI scans.

Segmentations of key anatomical structures, whether imported, manually created, or generated automatically using artificial intelligence (AI), are performed on these images, and together they form the final hologram (see **Figure 1**). These holograms are then displayed in the HL2 application, enabling surgeons to visualise the patient's internal anatomy and plan the surgery. While currently focused on preoperative planning, the system is being further developed for real-time guidance during EVD placement.

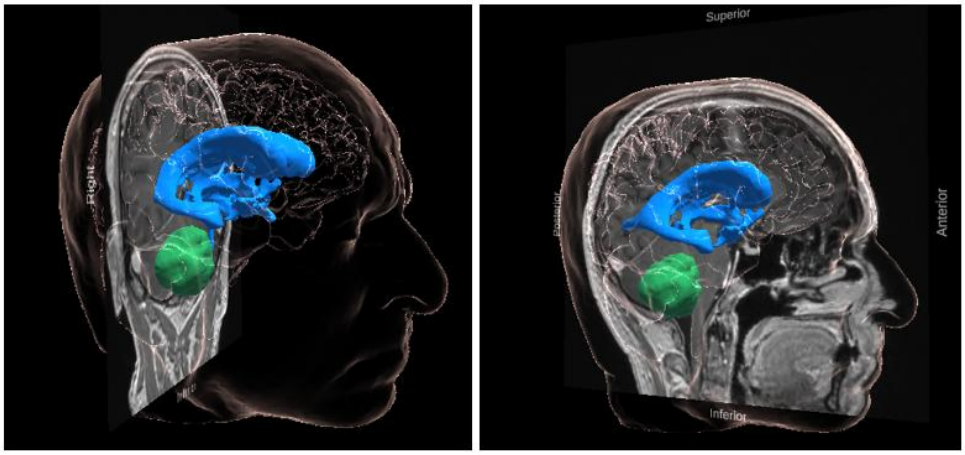

Figure 1: Example of a patient-specific hologram generated from MRI data in Lumi – *Ventricles are shown in blue, the tumour in green, and the skin and brain as a transparent surface.*

To bridge the gap between preoperative planning and intraoperative guidance, image-to-patient registration is essential. This registration process aligns the 3D patient model with the patient's physical anatomy (19). The Lumi software enables point-based registration using a head-mounted device and a pointer, both equipped with optical markers. This technique involves identifying corresponding points, such as fiducials or anatomical landmarks, on the 3D patient model and on the patient's head using the specialised pointer (19). Other prospective studies have evaluated Lumi's registration performance using MRI

with fiducials or CT with fiducials; however, these configurations do not reflect the workflow used for urgent EVD placement. In that setting, the most practical approach is to use anatomical landmarks together with the routine diagnostic CT. The accuracy achievable under this specific configuration has not yet been characterised. Fiducial marker registration would be more accurate, but it requires additional time for marker placement and imaging, making it less suitable for emergencies.

Currently, Lumi's gold standard for planning those landmarks on the virtual patient is manual annotation within the HL2 application. Users position landmarks by pinching and dragging arrows with their fingers, but this manual process can be time-consuming and cumbersome. To enhance efficiency, an AI-driven algorithm that automates the annotation of anatomical landmarks on holograms has been incorporated into Lumi. Using a dataset containing both MRI and CT images, de Boer et al. achieved a mean Euclidean distance of 4.01 mm (standard deviation (SD), 2.64 mm) with this algorithm (20). This means that the AI-predicted landmarks were, on average, 4 mm away from the reference (manual) landmark positions. Building on these initial results, the next step is to evaluate the algorithm's performance more thoroughly by focusing on clinicians' assessment of the AI-generated landmarks.

Therefore, this thesis aims to evaluate the clinical feasibility, robustness and registration accuracy of AR guidance using CT and anatomical landmarks for EVD placement, including an assessment of AI-assisted landmark annotation.

1.4 OBJECTIVE

The overarching objective of this thesis is to evaluate whether CT-based anatomical landmark registration using the Lumi AR workflow is sufficiently accurate, robust, and feasible to support EVD placement, thereby improving placement accuracy. This evaluation includes both the clinical suitability of AI-assisted anatomical landmark annotation and the performance of the complete AR-based registration workflow in the OR.

To address this objective, the following four subgoals are defined:

1. To assess the acceptability of AI-generated anatomical landmarks, focusing on perceived accuracy by clinicians, interobserver agreement, and the extent of manual adjustment required for clinical use.
2. To evaluate the registration accuracy of CT-based AR guidance using anatomical landmarks in the OR, with target registration error (TRE) as the primary outcome measure.
3. To examine workflow feasibility by quantifying the time required to perform AR-based registration within the clinical workflow and by collecting qualitative observations related to clinical integration.
4. To explore the robustness of the system by documenting technical stability, tracking reliability, and failures encountered during intraoperative use.

The work proceeds in two stages. An initial exploratory study evaluates the clinical suitability of AI-generated anatomical landmarks for AR-based registration during EVD placement. This is followed by a prospective pilot study that assesses the accuracy, usability, and robustness of the complete Lumi AR registration workflow in real patients, including an exploratory non-inferiority comparison with a prior phantom study.

The subsequent chapters follow the chronological order of the research process. Chapter II covers the study about AI landmark acceptability outside the OR, while Chapter III covers the OR registration study. Chapter IV presents the general discussion and conclusion, in which the findings from the preceding studies are integrated to address the goals of this thesis. Together, these studies form important preparatory steps towards implementing real-time AR-guided EVD placement in future clinical practice.

II. Clinical Acceptability of AI-generated Anatomical Landmarks

An essential part of image-to-patient registration in this thesis is the annotation of landmarks on the 3D patient model. Currently, the gold standard for landmark placement in Lumi is manual annotation using the HL2 application. This can be time-consuming and cumbersome. To reduce manual effort and errors during landmark placement, an AI solution has been developed. This could make AR navigation more practical and appealing for clinical implementation. During development, the algorithm's accuracy was quantified on a mixed dataset of CT and MRI scans, yielding a mean Euclidean distance of 4.01 mm (SD: 2.64 mm)(20). However, such theoretical measures do not necessarily reflect clinical relevance. In practice, no absolute ground truth exists: minor deviations in landmark placement are often acceptable as long as the landmarks can be reliably applied to the physical patient. This small, exploratory study, therefore, seeks to complement those quantitative results with a more practice-oriented assessment of how clinicians perceive the AI-generated landmarks. The findings serve as an initial exploration of the accuracy and reliability of AI-generated landmarks in the clinical context of EVD placement, thereby informing their suitability for the subsequent clinical registration study in the OR (see Chapter III).

2.1 METHODS

2.1.1 Ethics

The study was conducted under a non-WMO (Wet medisch-wetenschappelijk onderzoek met mensen) protocol approved by the Medical Ethics Review Committee (METC) of the University Medical Centre (UMC) Utrecht, which permitted the use of anonymised patient and imaging data for training and validation of AI algorithms. The METC approved a waiver of informed consent, given that the study involved a substantial amount of retrospective data. All data were handled in accordance with institutional regulations and anonymised before analysis to ensure patient confidentiality.

2.1.2 Data Source

CT scans of adult patients who underwent EVD placement at the UMC Utrecht between February and June 2025 were retrieved under the approved non-WMO protocol. Imaging data were exported from the Picture Archiving and Communication System (PACS), anonymised, and subsequently imported into the Lumi software (see **Appendix A** for software versions and build details).

2.1.3 Eligibility criteria

Inclusion criteria were:

- Adult patients (≥18 years) who underwent EVD placement.
- A field of view (FOV) that included both eyes and ears, ensuring all landmarks were visible for annotation.
- Thin-slice CT acquisition with slice thickness ≤ 1.0 mm and a matrix size of 512 × 512 pixels, required for accurate anatomical landmark placement.

Exclusion criteria were:

- CT acquisition issues, including motion artefacts, reconstruction artefacts, or other image-quality deficits that impaired reliable 3D skin-surface generation.
- Presence of external objects such as oxygen masks, fixation devices, dressings, or hardware that interfered with facial anatomy or segmentation.
- Segmentation-quality issues, including incomplete or distorted skin segmentation that failed visual quality control (e.g., missing eyes or ears, surface defects).

2.1.4 Data Preprocessing

Once imported into the Lumi cloud environment, skin segmentation was performed using the internally developed 'CT non contrast Cranial' algorithm integrated in the Lumi software. Each segmentation was visually reviewed according to the above quality criteria.

Anatomical landmark annotation

For all included scans, seven anatomical landmarks were automatically placed by the AI model: nasion, left and right medial canthi, left and right lateral canthi, and left and right auricular roots (located where the ear cartilage attaches to the skull) (see **Figure 2**). No manual corrections were made, as the goal was to evaluate the accuracy of the AI-generated placements.

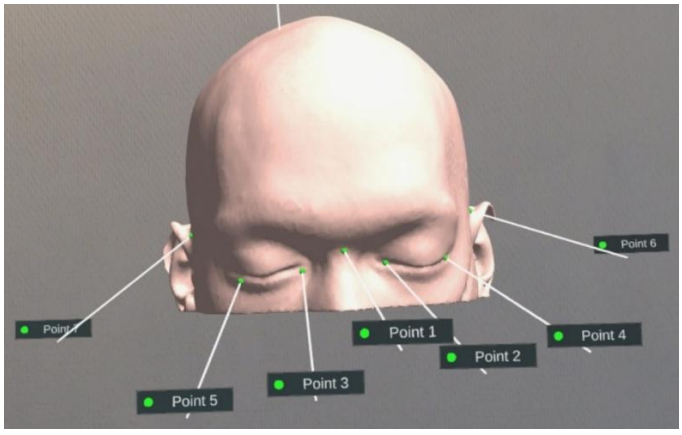

***Figure 2: Visualisation of the seven anatomical landmarks placed by the AI tool** – (1) nasion, (2–3) medial canthi, (4–5) lateral canthi, and (6–7) auricular roots.*

2.1.5 Experiment Setup

Four clinicians from the neurosurgery departments at UMC Utrecht and Amsterdam UMC were asked to assess the quality of the AI-generated landmarks. The estimated duration of each session was approximately 30 minutes per participant. Participants had prior experience with Lumi and AR, and only basic familiarity was required, as the study involved simply opening and viewing the hologram. Before starting, participants completed a questionnaire that captured their experience with point-based registration and AR, as well as their level of medical training. Experience was categorised as follows:

- No experience: 0–1 prior uses*
- Basic experience: 2 or more prior uses
- Experienced: at least monthly use for six months or more, currently or in the past

** This category was included for completeness, although no participants were expected to fall into it due to the requirement for basic familiarity with AR and Lumi.*

Participants were provided with an information sheet detailing the purpose and procedure of the study (see **Appendix B**). They then viewed the 15 holographic head models, each displaying AI-generated landmarks, in a randomised order. For each hologram, participants were asked:

"Which landmark would you adjust if you were in the OR and intended to perform a point-based registration? Consider the seven predefined landmark locations and ensure they can be accurately translated to the physical patient."

2.1.6 Outcome measures

The primary outcome of the study was the landmark adjustment rate, defined as the proportion of AI-generated landmarks that reviewers modified. The secondary outcome was interobserver agreement, assessed using unanimous agreement, partial agreement, and Fleiss' kappa. Unanimous agreement was defined as complete concordance among all observers, whereas partial agreement was defined as the percentage of observers who gave the most common rating.

2.1.7 Data Collection

All data were systematically logged in Microsoft Excel (v2510, Microsoft Corporation, Redmond, WA, USA) for subsequent analysis. For each participant, adjustments to the AI-generated landmarks were recorded as binary ratings (0 = Accept, 1 = Adjust). Free-text notes were also collected to provide qualitative context on the acceptability and usability of the landmarks. In addition, pre-experiment questionnaires collected information on participants' experience with point-based registration and AR, their level of medical training, and their affiliated hospital.

2.1.8 Data Analysis

Data analysis was performed using Microsoft Excel and Python (v3.10.9; Python Software Foundation, Wilmington, DE, USA). The analysis focused on two components: (1) the frequency of landmark adjustments and (2) interobserver agreement.

Adjustment rates were calculated overall and for each landmark, hologram, and observer. Because adjustment is a binary outcome (adjusted vs. not adjusted) and follows a binomial distribution, results were expressed as the proportion of landmarks adjusted out of all AI-generated landmarks with Wilson 95% confidence intervals (CIs).

Interobserver agreement was assessed using three metrics. Two of these were evaluated for each hologram–landmark combination and summarised overall, as well as stratified by landmark and by hologram:

I. **Partial agreement**, defined as the percentage of observers giving the most common rating (possible values: 50%, 75%, or 100%). Results were presented as median with interquartile range (IQR) because the metric is ordinal and discrete.
II. **Unanimous agreement**, defined as complete concordance among all observers (scored as 1 if unanimous, otherwise 0). As a binary measure, unanimous agreement was reported as proportions with 95% CIs.

Additionally, Fleiss' kappa was calculated to quantify overall agreement while accounting for chance.

In addition to the quantitative analysis, free-text notes were reviewed descriptively to identify recurring themes. All summary tables, including overall adjustment rates, per-landmark and per-hologram rates, and interobserver agreement measures, were exported to Microsoft Excel for visualisation and reporting.

2.2 RESULTS

2.2.1 Dataset Characteristics

A total of 23 patients with available CT scans were initially identified. Of these, nine were excluded: two due to head deformation, one due to motion artefacts, two due to interference from external objects affecting skin reconstruction, one due to an export failure, and three due to an incorrect FOV. This resulted in 15 patients meeting all inclusion criteria. The included cohort had a mean age of 66 years (range, 20–84 years) and consisted of 9 men and 6 women. The corresponding CT scans contained 130–285 slices, depending on the acquired FOV. The in-plane resolution ranged from 0.39 to 0.50 mm (mean: 0.44 × 0.44 mm), and the slice thickness was 0.9 or 1.0 mm for all scans.

2.2.2 Participant Characteristics

The four clinicians who participated included one non-specialist doctor (ANIOS) and three neurosurgical residents (AIOS). For both AR and point-based registrations, experience

levels were similar: one participant reported basic experience, whereas the other three were classified as experienced (see **Table 1**).

Table 1: *Baseline characteristics of the participants*

Characteristic	Number (%)
Participants	
Total	4 (100)
Affiliated hospital	
Amsterdam University Medical Centre	1 (25)
University Medical Centre Utrecht	3 (75)
Level of medical training	
ANIOS*	1 (25)
AIOS†	3 (75)
AR experience	
None‡	0 (0)
Basic§	1 (25)
Experienced‖	3 (75)
Point-based registration experience	
None‡	0 (0)
Basic§	1 (25)
Experienced‖	3 (75)

Abbreviations: *AR = Augmented Reality*
** ANIOS = non-specialist doctor (Arts Niet In Opleiding tot Specialist)*
† AIOS = resident (Arts In Opleiding tot Specialist)
‡ None = zero or one prior uses
§ Basic = two or more prior uses
‖ Experienced = at least monthly use for six months or more, currently or in the past

2.2.3 Primary Outcome: Adjustment Rate

In total, the 15 holograms, each containing 7 anatomical landmarks, assessed by 4 clinicians, resulted in 420 data points. The overall adjustment rate across all landmarks, holograms and participants was 22.9% (95% CI, 19.1–27.1%). **Table 2** presents the adjustment rates by participant, hologram, and anatomical location.

Table 2: *Adjustment rates of AI-generated anatomical landmarks by participant, hologram and anatomical location*

ID	Adjusted landmarks n (%)	95% CI (%)
Participant*		
A	27 (25.7)	18.3–34.8
B	28 (26.7)	19.1–35.8
C	33 (31.4)	23.3–40.8
D	8 (7.6)	3.9–14.3
Hologram†		
1	6 (21.4)	10.2–39.5
2	6 (21.4)	10.2–39.5
3	3 (10.7)	3.7–27.2
4	4 (14.3)	5.7–31.5
5	1 (3.6)	0.6–17.7
6	5 (17.9)	7.9–35.6
7	4 (14.3)	5.7–31.5
8	11 (39.3)	23.6–57.6
9	16 (57.1)	39.1–73.5
10	6 (21.4)	10.2–39.5
11	3 (10.7)	3.7–27.2
12	9 (32.1)	17.9–50.7
13	0 (0.0)	0.0–12.1
14	11 (39.3)	23.6–57.6
15	11 (39.3)	23.6–57.6
Anatomical location‡		
1 – nasion	1 (1.7)	0.3–8.9
2 – medial canthus (l)	3 (5.0)	1.7–13.7
3 – medial canthus (r)	15 (25.0)	15.8–37.2
4 – lateral cantus (l)	19 (31.7)	21.3–44.2
5 – lateral cantus (r)	23 (38.3)	27.1–51.0
6 – auricular root (l)	22 (36.7)	25.6–49.3
7 – auricular root (r)	13 (21.7)	13.1–33.6

Abbreviations: *n = number; CI = confidence interval; l = left; r = right.*

** Each participant rated a total of 105 landmarks (15 holograms × 7 landmarks)*

† Each hologram received 28 ratings (4 participants × 7 landmarks)

‡ Each anatomical location received 60 ratings (4 participants × 15 holograms)

Adjustment rates differed across participants. Participant D had the lowest rate at 7.6%, whereas participants A, B, and C had higher and relatively similar rates of 25-31%. Excluding participant D, who was an outlier, the average adjustment rate among the remaining participants (A–C) was 27.9% (95% CI, 23.3–33.1). Adjustment rates varied across the 15 holograms, ranging from 0.0% (hologram 13) to 57.1% (hologram 9). While most holograms exhibited rates below 25%, several (holograms 8, 9, 12, 14, and 15) showed higher adjustment rates above 30%. Among the anatomical landmarks, the lateral canthi and left auricular root were most frequently adjusted (>30%), while the nasion was rarely adjusted (1.7%).

2.2.4 Interobserver Agreement

The overall interobserver agreement across all landmarks and holograms had a median partial agreement of 100.0% (IQR, 25.0%). Unanimous agreement had a mean of 61.0% (95% CI, 51.4–69.7%). Fleiss' kappa yielded a value of 0.42, indicating moderate agreement.

By hologram, partial agreement was generally high, with most median values reaching 100.0%. Unanimous agreement per hologram was more variable, ranging from 28.6% to 100.0%, with most values exceeding 50%. Hologram 13 achieved unanimous agreement across all observers and landmarks, whereas holograms 8, 9, and 12 had the lowest median partial agreement (75.0%) and the lowest unanimous agreement (28.6%). When considering both partial and unanimous agreement, the nasion scored highest (100.0% and 93.3%, respectively), while the right medial canthus, left lateral canthus, and left auricular root scored lowest (all 75.0% and 46.7%, respectively). **Appendix C** provides the entire table with the results for partial and unanimous agreement.

Figure 3 presents the adjustment rate and unanimous interobserver agreement for each of the seven anatomical landmarks. Landmarks with lower adjustment rates generally showed higher interobserver agreement. The nasion in particular stands out, showing both a relatively low adjustment rate and high interobserver agreement.

2.2.5 Qualitative Observations

Across participants, a consistent observation was that, although no landmark was entirely misplaced relative to its intended position, many were slightly offset, typically by 1–3 mm. While these deviations were generally minor, they were consistently mentioned as a limitation for accurate registration in emergency settings. Still, participants agreed that with careful inspection and sufficient time, the landmarks could often be interpreted and transferred to the corresponding locations on a physical patient. Participant D particularly emphasised this point.

Several clinicians noted that the auricular root identified by the algorithm was less familiar to them in clinical practice, as they typically use the tragus as a landmark in this region. Participant A observed that landmarks, intended to be positioned at the nasion, were frequently located closer to the glabella. Although this deviation will not necessarily hinder registration, it is inconsistent with the algorithm's intended definition of the nasion. Finally,

Participant C highlighted variability in the positioning of the lateral canthus, which was sometimes placed directly on the orbital rim and other times more medially. Participant C considered the latter positioning less desirable, as it corresponds to a non-rigid region near the eyeball.

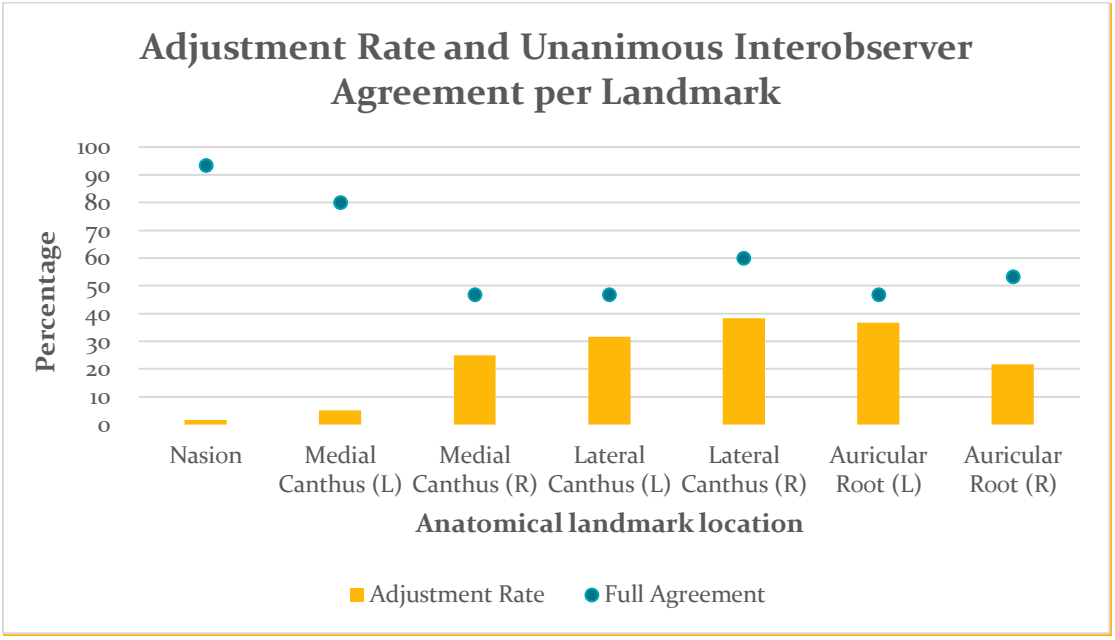

Figure 3: Overview of the adjustment rate and unanimous interobserver agreement for all seven anatomical landmark locations *– L = left; R = right.*

2.3 DISCUSSION

This exploratory study investigated the clinical acceptability of AI-generated anatomical landmarks in Lumi and quantified clinicians' agreement on the need for manual adjustments. These evaluations contributed to the broader aim of determining whether CT-based anatomical landmark registration using the Lumi AR workflow is sufficiently accurate, robust, and feasible to support and eventually improve EVD placement. Overall, the results indicated that the AI algorithm achieves a high degree of accuracy, with an average adjustment rate of only 22.9% (95% CI, 19.1-27.1%) and a median partial interobserver agreement of 100.0% (IQR, 25.0%). However, unanimous agreement was lower at 61.0% (95% CI, 51.4–69.7%), and Fleiss' kappa (0.42) indicated only moderate overall consistency between raters. Although 95% CIs and IQRs were calculated, the variability they reflect was expected mainly due to the small sample size and the inherently subjective nature of the experiment. The nasion was the landmark most consistently placed correctly, requiring few adjustments, and showed strong consensus among clinicians.

During the course of the study, a consistent pattern in the first participants' responses became evident. The experiment was therefore concluded after four participants instead of seven, and their data were included in the analysis. As a result, intra-rater testing was also

not performed, since additional sessions were unlikely to yield new insights. Around the same time, the AI functionality was undergoing internal testing for a new workflow implementation, and feedback from those evaluations supported the decision to end the experiment early.

Participant-related factors were considered as potential sources of variability. All participants had comparable AR/Lumi experience, and none of the less-experienced users behaved as outliers. As such, differences in AR proficiency are unlikely to have influenced the findings. Direct viewing and interaction with the holograms in the HL2 headset were considered essential for providing true three-dimensional spatial perception and depth cues, which cannot be replicated on conventional displays. The required AR skills were minimal, and only participants with prior HL2 experience were included to ensure familiarity with basic operations, such as opening and manipulating holograms.

When considered individually, landmarks showed patterns that explained variability in clinician agreement. Rigid, well-defined bony structures, such as the nasion, were easiest to identify and had the highest interobserver agreement and lowest adjustment rate. In contrast, landmarks around the eyes, particularly the lateral canthi, were less discrete and relied on interpretation of soft-tissue contours. This resulted in greater observer dependence and higher adjustment rates. The auricular roots also showed increased variability, possibly due to the gradual transition from the skin covering the cartilage to the scalp in the surface model. With such a transition area, precise point definition is complicated without tactile feedback. Moreover, the reliability of landmark placement is inherently linked to the accuracy of the skin segmentation.

Taken together with the qualitative feedback, these findings suggested that, while the algorithm generally produces accurate and acceptable landmark placements, subtle deviations of 1–3 mm occur. These deviations were relatively minor and consistent with findings from the algorithm developers' initial testing during development. However, clinicians noted that they could be significant in high-stakes, time-sensitive scenarios, such as emergency EVD placement (20). Thus, despite acceptable overall accuracy metrics, the algorithm in its current form is not yet suitable for reliable use during EVD placement and will therefore not be used in the subsequent clinical registration study.

2.3.1 Limitations

Several limitations should also be acknowledged. First, this was a small-scale, subjective study with a limited number of participants, and the ratings inherently reflect personal interpretation rather than an actual objective ground truth. Second, the assessment focused solely on visual inspection of landmarks in AR rather than actual registration performance in the OR using those landmarks. Third, the evaluation was performed using only the holographic models, without direct comparison with the real patients. Consequently, factors that may influence landmark placement in real-world conditions, such as lighting, patient positioning, and soft-tissue deformation, were not accounted for. Annotating the

landmarks on the hologram while simultaneously viewing the physical patient might be a more optimal approach. However, this approach was not feasible in the current study due to the dataset's retrospective design and is not suitable in all situations, as it may be preferable to complete preparations outside the OR.

2.3.2 Clinical Feasibility

From a clinical perspective, AI-generated landmarks appear promising but are not yet suitable for emergency procedures like EVD placement, where accuracy and speed are critical. Occasional corrections and the need for careful verification reduce the potential time savings. Using the AI landmarks requires deliberate inspection, which may be impractical in acute scenarios. This study did not involve a physical patient, but the findings remain relevant: landmarks that showed high variability in a controlled virtual environment may be even more challenging to consistently identify during real-world registration. In addition, informal feedback from neurosurgeons highlighted limitations of the HL2 interface: selecting and adjusting landmarks with hand gestures can be imprecise and occasionally cause unintended movements or deletions, making fine adjustments time-consuming. Nevertheless, the tool will perform well in less urgent, controlled settings, where there is sufficient time to review and adjust landmark positions.

2.3.3 Future Directions

First, the variability observed across anatomical landmarks highlights opportunities for improvement. Landmarks with low interobserver agreement, such as the lateral canthi and auricular roots, could be refined by clarifying their precise definitions to improve consensus, or potentially replaced with alternative points. However, the total number of suitable landmarks on the head is limited. In contrast, the nasion demonstrated high consistency and could serve as a reliable validation point in future workflows.

For future development, transitioning the landmark placement and review process from the HL2 environment to a web-based interface could improve usability. Prototypes of such interfaces have already been tested and show potential to make landmark placement faster and more intuitive. Additionally, retraining or fine-tuning the algorithm using clinician feedback could help reduce systematic errors. Ultimately, combining AI-generated landmark placement with a web-based interface would enable automatic landmark suggestions and easy adjustments within the same platform. This, however, represents a longer-term objective requiring larger datasets and iterative validation in clinical settings.

2.4 CONCLUSION

This study explored AI-generated annotation of anatomical landmarks to develop a workflow suitable for emergency settings, such as EVD placement. While the current algorithm demonstrated generally accurate placements (overall adjustment rate: 22.9%), subtle deviations and the need for careful verification limited its readiness for high-stakes, time-critical scenarios. The findings highlighted the key challenges – speed, accuracy, and

intuitive interaction – that must be addressed to develop a clinically viable tool for emergency use. Consequently, AI-generated landmarks were deemed not to meet the requirements and were therefore excluded from further clinical testing.

In the short term, efforts should focus on improving usability by enabling manual landmark placement within the Lumi web application. Long-term development should aim to optimise the AI algorithm's accuracy further using clinician feedback and larger datasets.

III. Registration Accuracy in the Operating Room

In the previous chapter, the clinical suitability of AI-generated anatomical landmarks was evaluated. Although these landmarks were designed to enable automatic annotation on a 3D patient model and thereby support image-to-patient registration, they were not yet considered sufficiently reliable for direct clinical use. However, landmark annotation accuracy alone does not fully determine the performance of an image-to-patient registration. Instead, registration accuracy is the result of the complete workflow, including image acquisition, hologram generation, landmark selection, and registration execution.

A widely used method to quantify the accuracy of image-to-patient registration is to measure the TRE. TRE provides an independent, clinically relevant measure of how well the virtual model aligns with the patient in physical space (21). Typically, sub-2 mm accuracy is required for many neurosurgical interventions (22, 23). However, a TRE of approximately 5 mm is considered clinically acceptable for EVD placement in this pilot study, given the relatively large size of the ventricular system.

To date, the registration accuracy of the complete Lumi AR workflow has not yet been evaluated in the OR, despite this being a critical step towards clinical implementation. An earlier phantom study demonstrated the technical feasibility and potential clinical value of AR-assisted EVD placement using point-based registration with anatomical landmarks and CT-derived holograms. However, validation in a real clinical setting remains necessary (24).

Accordingly, this chapter shifts the focus from individual components to an evaluation of the complete AR-based registration process on patients. This research can be placed within the IDEAL framework: a structured approach for assessing innovative surgical technologies across five successive stages of development (**I**dea, **D**evelopment, **E**xploration, **A**ssessment, **L**ong-term follow-up). The work presented in this thesis corresponds to Stage 2a (Development) of this framework, as it involves iterative refinement toward a stable system and feasibility. The focus is on validating a specific technical component of this new EVD workflow - AR-based registration of the 3D model to the patient - rather than evaluating the complete surgical procedure (25).

The primary aim of the pilot study is therefore to prospectively assess the registration accuracy, feasibility and system robustness of the Lumi AR workflow in the OR using CT-based holograms and manually annotated anatomical landmarks. Eventually, this would contribute to an improved drain placement accuracy.

3.1 METHODS

3.1.1 Study design

This study consisted of two sequential phases: a formative phase and a summative phase. Both phases evaluated point-based image-to-patient registration using the Lumi AR system in a clinical environment. The formative phase focused on iterative testing and refinement of the software, whereas the summative phase was designed as a prospective pilot study to determine whether the workflow demonstrated sufficient accuracy, feasibility, and robustness to justify further clinical implementation (26).

Overall workflow

The complete workflow comprised preparation steps performed outside the OR and registration steps performed inside the OR. Outside the OR, a CT-derived 3D patient model was created, and anatomical landmarks were annotated using the Lumi web application. Inside the OR, the patient was registered to the virtual model using point-based registration with anatomical landmarks via the Lumi HL2 application. Registration accuracy was then assessed using predefined validation points. This workflow was identical in both phases, although only data from the summative phase were included in the primary analysis.

Formative Phase

The purpose of the formative phase was to refine the AR-based registration workflow and ensure technical stability and feasibility before formal summative evaluation. Formative testing was conducted at UMC Utrecht across multiple patients. No predefined sample size was set. Both qualitative feedback and quantitative registration metrics were collected iteratively and communicated to the development team. This process continued until no new critical issues were encountered, at which point the summative pilot study was initiated. Data from this phase were not included in the final accuracy analysis.

Summative Phase (pilot study)

This phase was designed as a prospective pilot study to assess the system's registration performance in real clinical practice. The pilot study was conducted at the UMC Utrecht, and patients were included between November 24, 2025 and December 2, 2025. The primary objective was to evaluate the registration accuracy of the Lumi AR tool for point-based registration of CT-derived 3D patient models using manually placed anatomical landmarks. Approximately 10-15 patients were included.

3.1.2 Eligibility criteria

Adult patients (≥18 years) admitted to the Department of Neurosurgery at UMC Utrecht and scheduled for cranial surgery under full sedation were included. Inclusion required a preoperative CT scan obtained within the previous six months, with no history of cranial surgery or major physical changes affecting facial anatomy since the scan. To ensure

accurate landmark annotation, the CT slice thickness had to be ≤ 1.0 mm, and the FOV had to encompass at least both orbits and the external auditory meatus.

3.1.3 Ethics

Patient recruitment was conducted under ethical approval from the UMC Utrecht METC. The study was classified as an amendment to a previously approved non-WMO protocol for an MRI- and fiducial-based registration study. All participants received verbal and written information about the study, and written informed consent was obtained before inclusion. At the time of the study, the Lumi software was under active development and in progress toward CE (Conformité Européenne) certification. All data were handled in accordance with institutional regulations and anonymised before analysis to ensure patient confidentiality.

3.1.4 Materials

The following equipment and software were used:

- The HL2, which was used to visualize the 3D patient models in AR (see **Figure 4**).
- The Lumi software, which consists of two components:
    - Lumi web application, integrated within the UMC Utrecht PACS infrastructure, for hologram creation.
    - LumiNE Elite module for the HL2 for registration (referred to as the Lumi HL2 application).

  Detailed software versions and build information are provided in **Appendix A**.
- Custom registration tools, which included a stainless steel head-mounted reference device and a pointer, both equipped with engraved optical markers (Vuforia, PTC, Boston, USA). The head-mounted device was secured to the patient's forehead and nose, and served as a stable tracking reference (See **Figure 5**)(24).

Figure 4: Microsoft HoloLens 2 *(27)*

Figure 5: Hardware toolset for AR registration *– From left to right: front view of the head device with reference marker, side view of the head device with reference marker, and the pointer.*

3.1.5 Detailed workflow steps

Figure 6 illustrates the complete study workflow. The individual steps are described in detail below. The step labelled "EVD workflow" refers specifically to the sequence of actions within the Lumi HL2 application used to perform point-based registration; these steps are presented in the same order and using the same terminology as in the application. All procedures were performed by a single researcher (MG), who had prior experience with over 30 point-based registrations using Lumi.

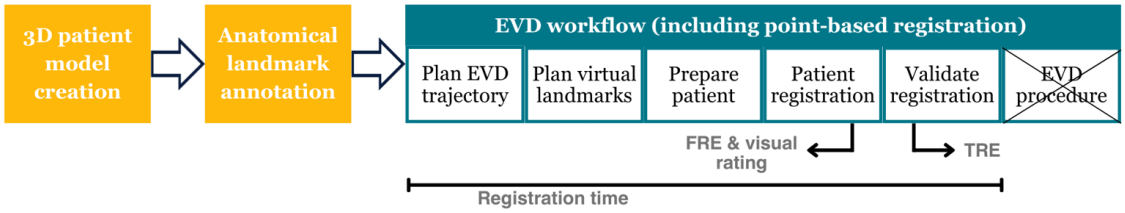

Figure 6: Schematic overview of the study workflow *– Yellow indicates steps performed outside the operating room in the Lumi web application, and blue indicates steps performed in the operating room using the Lumi HL2 application. The arrows/lines indicate at which step each outcome was collected. The actual EVD procedure was not performed in this study. EVD = external ventricular drain; FRE = fiducial registration error; TRE = target registration error.*

3D Patient Model Creation

Imaging data were exported from PACS, anonymised, and imported into Lumi. For each patient, a 3D stereoscopic model was generated from a preoperative CT scan using the Lumi web application. The integrated CT non-contrast Cranial Segmentation Function (CTncSF; Augmedit, Naarden, The Netherlands) within Lumi automatically segmented the skin, skull, brain, and ventricles. CTncSF is a deep learning algorithm based on nnU-Net and trained on manually annotated datasets. It produces 3D segmentations, which are exported as meshes for visualisation in AR. Manual skin segmentation within Lumi was used when fiducial markers were not included in the automatic skin segmentation.

Anatomical Landmark Annotation

Since earlier analyses (see Chapter II) indicated that automated landmark annotation was not yet sufficiently accurate for EVD scenarios, and manual landmark placement in the Lumi web application showed potential, a manual annotation function was implemented in the web application and used for this study. Six registration points and one to three validation points were manually placed for each patient with this new software feature.

The preferred landmarks as registration points were the left and right auricular roots (cartilaginous ear–skull junction), the left and right lateral canthi, the right medial canthus, and the subnasale (28). The subnasale was preferred over the left medial canthus to achieve a better spatial spread. When the scan's FOV did not include the area below the nose, the

left medial canthus was annotated instead of the subnasale to maintain a consistent number of registration points.

For validation, one to three additional points were annotated: the nasion (included for all patients based on the results in Chapter II) and, when available, the two fiducial markers closest to Kocher's point, the usual entry site for EVDs (see **Figure 7**). Fiducial markers were present only in cases using neuronavigation and were used exclusively for TRE evaluation, not for registration. Fiducials were preferred for validation because they can be identified unambiguously in both physical and virtual space and can be positioned near the intended drain trajectory, where reliable anatomical landmarks are typically absent.

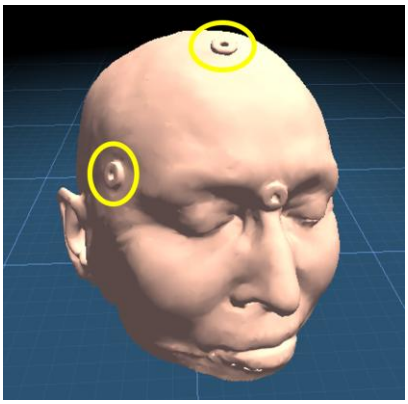

Figure 7: Example of skin model with fiducial markers *– The yellow circles indicate the two fiducial markers closest to Kocher's point.*

Point-Based Registration

Point-based registration was performed in the OR prior to fixation of the Mayfield head clamp, simulating the clinical EVD placement conditions, which is conducted without rigid head fixation. All patients were positioned supine on a horseshoe headrest. The main steps of the registration workflow within the HL2 application that were used in this study were:

- **Plan landmarks:** During this step, the researcher verified that the landmarks placed during preparation outside the OR were correctly positioned.
- **Prepare patient:** The incision-planning component of this step (used clinically) was omitted. The head-mounted reference device carrying an optical marker was positioned on the patient's forehead and secured with surgical tape (see **Figure 8**). Tracking was then activated, and the researcher verified the reference marker by fixating on it for several seconds. If the detected outline was inaccurate, the calibration option was used.
- **Patient registration:** Using the tracked pointer, each annotated landmark was indicated on the physical patient, after which the registration was computed (see **Figure 8** & **Figure 9**). If the resulting registration was unsatisfactory, one or more poorly indicated landmarks could be re-annotated, and the registration repeated. Once registration was complete, the researcher assessed visual alignment by

observing the hologram outline of the skin through the HL2 as the researcher moved around the patient, focusing on key facial landmarks.

- **Validate registration:** One or three validation points (depending on the available landmarks) were indicated on the physical patient to compute the TRE.

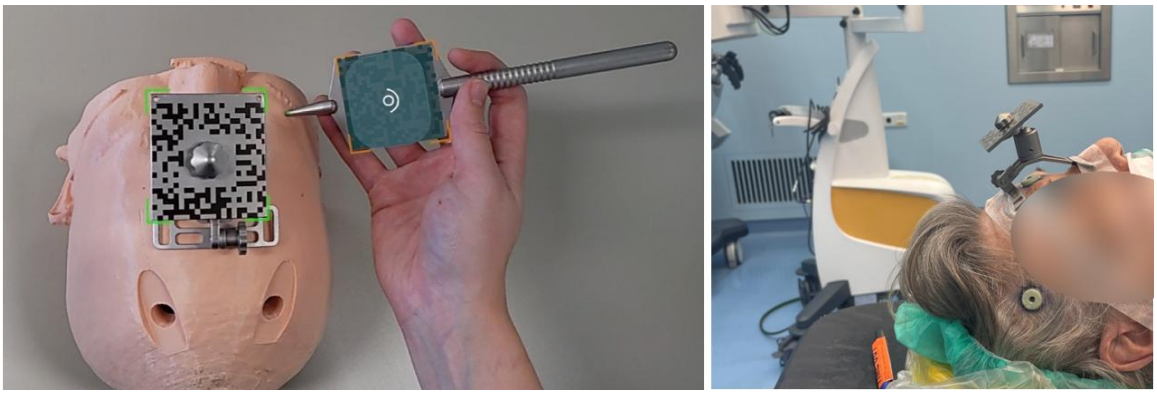

Figure 8: Point-based registration using the HoloLens 2 *– View through the HoloLens 2 during point-based registration on an example phantom (left) and positioning of the head-mounted device on a patient (right).*

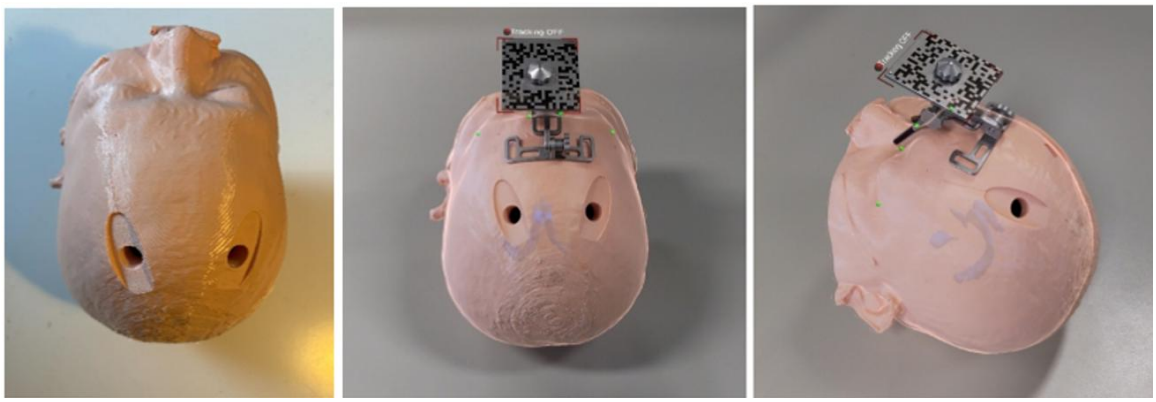

Figure 9: Holographic projection registered on a head phantom *– The bare phantom (left) and an example of a holographic projection on the phantom (middle and right). The glowing line on the skin indicates the skin model registration, and the transparent blue structures within the head represent the registered ventricles.*

3.1.6 Outcome Measures

Primary Outcome

Target Registration Error (TRE) – TRE was the primary outcome measure, as it directly reflects clinically relevant registration accuracy. It is defined as the Euclidean distance between a virtual target point and its corresponding physical point that was not used for registration, providing an independent measure of alignment accuracy (21).

Secondary Outcomes

The secondary outcomes were defined as follows:

Fiducial Registration Error (FRE) - FRE quantifies global registration fit and is defined as the root mean square of the localisation error at the registration landmarks. It was included as a secondary outcome because it describes overall alignment quality, but does not directly reflect target-level clinical accuracy (21).

Visual Registration Accuracy Rating - A 5-point Likert-scale assessment (1 = very poor, 5 = excellent) of overall registration alignment based on visual inspection of the hologram outline at key facial landmarks (tip of the nose, ears, and back of the head). The assessment was performed by the same researcher who conducted the registration procedure.

Registration Time - Time from initiation of the EVD workflow in the Lumi HL2 application until all registration and validation points are placed. This should be considered an approximate measure, intended to provide a general sense of workflow speed.

System robustness – Documentation of technical stability during registration in the OR. Events were classified by their potential impact: critical failures were defined as events that could prevent workflow completion in a real clinical scenario, whereas recoverable failures disrupted the workflow but could be resolved without losing the registration.

Workflow feasibility - Assessment of practical aspects during registration in the OR, including ease of use, integration with standard clinical workflow, environmental factors (e.g., lighting, reflections), etc.

3.1.7 Data Collection

The Lumi software automatically computed the 3D (x, y, z) offsets between corresponding virtual and physical landmarks for both FRE and TRE calculations. These data were exported as structured Excel files. Registration time was also automatically saved by the software and displayed in a dashboard. Visual accuracy ratings and qualitative observations were documented in structured notes during or immediately after each procedure. All data were organised and compiled in Microsoft Excel for subsequent analysis.

3.1.8 Data Analysis

Data analysis was performed using Microsoft Excel and Python. Descriptive statistics were used to summarise registration metrics (FRE and TRE) and registration time, reported as mean (SD) and median (IQR). Visual accuracy ratings, as an ordinal outcome, were summarised using median (IQR) only. Qualitative data from feasibility assessments were analysed thematically to identify common challenges, benefits, and areas for improvement.

Exploratory non-inferiority analysis

Given the limited sample size, a definitive non-inferiority analysis was not feasible. An exploratory non-inferiority comparison was therefore performed using a predefined margin based on the prior phantom study. Because the phantom study reported distance-to-target of the drain tip rather than TRE, only FRE could be used as the metric for comparison (24).

Non-inferiority was tested using the following hypotheses:

- $\boldsymbol{H_0}$: $\bar{\mu}_{OR} - \bar{\mu}_{phantom} \geq \Delta$ (OR registration is inferior),
- $\boldsymbol{H_1}$: $\bar{\mu}_{OR} - \bar{\mu}_{phantom} < \Delta$ (OR registration is non-inferior),

Non-inferiority was defined relative to the phantom study mean FRE of 4.00 mm, using a predefined margin of 20% (Δ = 0.80 mm) (24). This margin was selected based on consensus with experienced clinicians. The OR registration was considered non-inferior if the upper bound of the two-sided 90% CI for the mean difference ($\bar{\mu}_{OR} - \bar{\mu}_{phantom}$) is below 0.8 mm.

For sample-size planning, a weighted two-sample z-based approach accounting for the phantom study variance indicated that 37 participants would be needed. Given the small pilot sample size (10–15 patients), two-sided 90% confidence intervals were calculated using the t-distribution in this thesis, and results were interpreted as exploratory. The complete derivation and rationale are given in **Appendix D**.

3.2 RESULTS

3.2.1 Formative phase

A total of five measurements were performed on different patients in the OR as part of the formative phase. The cohort consisted of 3 women and 2 men, with a mean age of 59.4 years (range, 40-79 years). In three cases, the FOV was sufficient to use the subnasale as a landmark; in the remaining two, the left medial canthus was used. In two procedures, the patient had undergone a neuronavigation CT scan instead of a standard CT scan, enabling the use of fiducials for validation. Surgical indications included tumour or metastasis resection (n=3), cerebral bypass surgery (n=1), and aneurysm clipping (n=1). Workflow duration was recorded for all measurements, with a mean of 5 min 12 s (SD, 1 min 18 s) and a median of 5 min 0 s (IQR, 2 min 48 s).

During the formative phase, several themes emerged regarding system performance and workflow. Overall, the registration was functional but showed occasional instability, including drift of the AR projection on the reference marker, shaky visualisations, and intermittent dropouts of the AR projections. For example, in one instance, even after registration was completed, the AR projection on the reference marker drifted by approximately 2 cm and remained offset. Environmental factors, such as patient repositioning and movement by surrounding staff or equipment, may have contributed to this variability. However, the system sometimes remained stable despite such activity. Despite these challenges, the registration process was generally considered smooth and

responsive, with minimal need for recalibration. Minor errors in landmark placement were observed, often related to inexperience with the workflow rather than software failure. FRE, TRE, and visual rating data were collected but were affected by these instabilities; they were used solely to guide iterative workflow refinement and are provided in full in **Appendix E**. Findings from the formative phase informed several modifications to the Lumi software, which are presented in **Appendix A**.

3.2.2 Summative phase

A total of 11 measurements were performed in the pilot study, all of which were successful. The cohort included 6 women and 5 men, with a mean age of 54.8 years (range, 19-86 years). In five cases, the FOV was sufficient to use the subnasale as a landmark; in the remaining six, the left medial canthus was used. Only four participants had undergone a neuronavigation CT scan, allowing the use of fiducials as validation points. However, in one of these cases, inaccurate skin segmentation prevented annotation of all fiducials, leaving only a single fiducial available for analysis. Surgical indications included tumour or metastasis resection (n=4), cyst resection (n=1), nerve decompression (n=3), pituitary surgery (n=2), and shunt placement (n=1).

3.2.2.1 Primary outcome

Of the 11 registrations performed, 10 included one or more TRE measurements. One TRE-nasion value had to be excluded because the software failed to save it correctly, and one measurement at fiducial1 failed due to inaccurate skin segmentation. **Table 3** provides an overview of the results, with more details presented in **Appendix E**.

Table 3: *Registration error metrics collected during the summative phase.*

Metric	Number of measurements	Mean (SD) (mm)	Median (IQR) (mm)
FRE	11	4.0 (0.7)	3.6 (1.2)
TRE-nasion	10	4.9 (2.1)	4.8 (2.0)
TRE-fiducial1*	3	7.4 (1.7)	6.9 (1.6)
TRE-fiducial2†	4	4.9 (1.9)	4.8 (2.6)

Abbreviations*: SD = standard deviation; IQR = interquartile range; mm = millimetres; FRE = fiducial registration error; TRE = target registration error.*
** Fiducial 1: Located along the midline near Kocher's point.*
† Fiducial 2: Located on the right side of the head near Kocher's point.

TRE values varied across the three locations, with the lowest errors at the nasion and fiducial2 and the highest at fiducial1. **Figure 10** provides an overview of TRE values across locations, showing that the values for the three TRE types vary even within a single patient case. Most of the time, fiducial1 (midline, top of head) had higher TRE values than the nasion, whereas fiducial2 (side of head) had lower TRE values than the nasion.

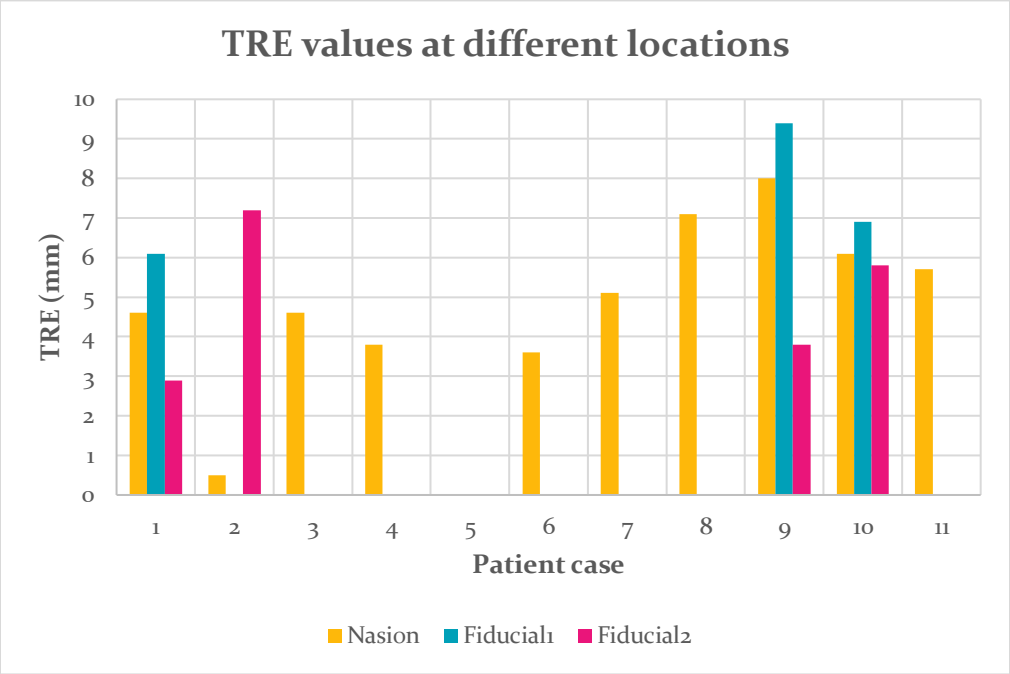

Figure 10: TRE values at the nasion and fiducial points per patient case during the summative phase *– Fiducial1 was placed in the midline on the superior part of the head, and fiducial2 was positioned laterally, both in proximity to Kocher's point. Patient cases 2 and 5 are missing data due to software issues. TRE = target registration error; mm = millimetres.*

3.2.2.2 Secondary outcomes

Fiducial registration error

The FRE for point-based registration in the OR had a mean of 4.0 mm (SD, 0.7 mm) and a median of 3.6 mm (IQR, 1.2 mm) across the 11 measurements (see **Table 3**).

Exploratory non-inferiority analysis

The difference in mean FRE between the OR and phantom study was –0.02 mm, with the upper bound of the two-sided 90% CI at 0.56 mm. Since this is below the pre-defined non-inferiority margin of 0.80 mm, OR registration is considered non-inferior to phantom registration. More details can be found in **Appendix E**.

Visual registration accuracy rating

Visual accuracy ratings across landmarks are presented in **Table 4**. Median scores were 5 (IQR, 1) for the nose, ears, and back of the head, indicating generally high perceived alignment. However, ratings varied between patients, with scores ranging from 3 to 5 across landmarks. The lowest ratings (score of 3) were observed at the nose and ears, whereas ratings for the back of the head did not fall below 4. Details can be found in **Appendix E**.

Table 4: *Visual registration accuracy ratings collected during the summative phase.*

Location	Number of measurements	Median (IQR)*	Range
Nose	11	5 (1)	3 - 5
Ear (right)	11	5 (1)	3 - 5
Ear (left)	11	5 (1)	3 - 5
Back of the head	11	5 (1)	4 - 5
Overall	**44**	**5 (1)**	**3 - 5**

Abbreviations*: IQR = interquartile range.*
** Scored on a 5 point Likert-scale (1 = very poor, 5 = excellent).*

Lastly, **Figure 11** presents TRE values alongside visual accuracy ratings for each measurement. TRE showed variability between measurements, whereas visual ratings were consistently high. No clear relationship was observed between the TRE and visual ratings.

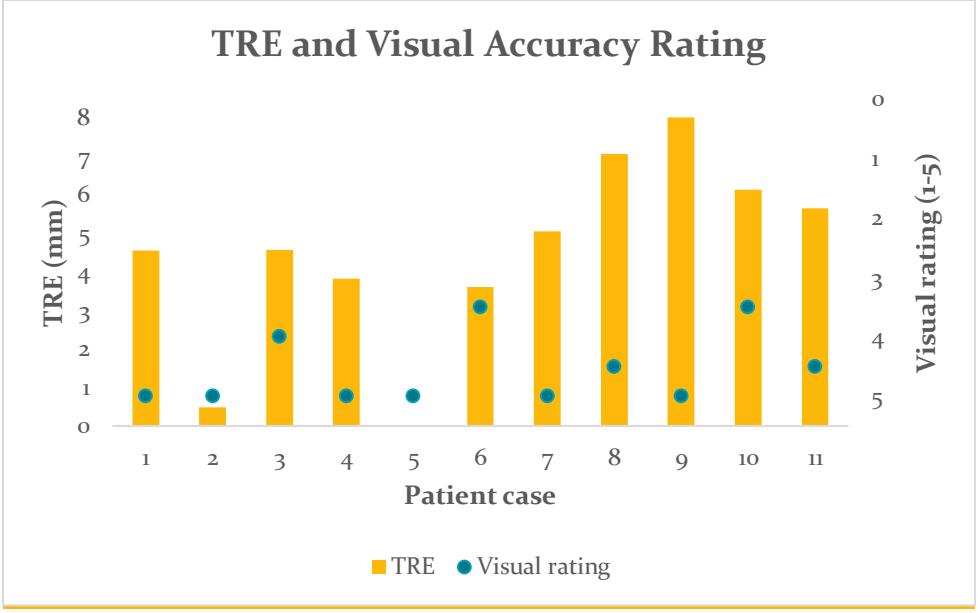

Figure 11: TRE and visual accuracy ratings across measurements during the summative phase *– Visual rating is the mean score across the four key facial landmarks and is scored on a 5-point Likert scale (1 = very poor, 5 = excellent). Patient case 5 misses the TRE metric due to a software problem. TRE = target registration error; mm = millimetres.*

Registration time

The total workflow time had a mean of approximately 4 min 36 s (SD, 1 min 12 s) and a median of 5 min 6 s (IQR, 1 min 48 s) across the 11 measurements. The duration per workflow step was also collected, but was not included in the primary analysis because execution of individual steps was inconsistent across measurements. These detailed timings are provided in **Appendix E**.

System robustness

During the measurements, several system instabilities were observed. Brief dropouts of the holographic visualisations and disruptive flickering of menus both occurred twice. Full software crashes also occurred twice. These events typically followed a pattern: while placing physical landmarks, an "environment unstable" message appeared along with a warning panel from Lumi (n=3). Afterwards, the software either crashed (n=2) or the menus were displaced in the room (n=1). Once the application was restarted or the menus repositioned, measurements could continue successfully. In another instance, the "environment unstable" message appeared without any near-crash, but in a measurement that was already generally feeling unstable. A similar feeling of instability was observed in another measurement. In both cases, the registration could still be completed successfully. On one occasion, instability was severe enough that continuation would have been difficult in a real clinical scenario. However, results were still obtained in this pilot.

Out of 10 instability events observed, 3 were classified as critical failures: 2 full software crashes and 1 severe instability that would have prevented continuation in clinical practice. In contrast, the remaining events (displacements, dropouts, and "environment unstable" warnings) were classified as recoverable failures. For all events that needed no restarting, any additional time due to instability was included in the registration time reported above. When a restart was necessary, which was only in case of a crash, this added up to five minutes to the procedure, and the registration timing restarted from zero.

Workflow challenges

Several practical challenges affected the registration workflow. Lighting varied across measurements, with insufficient light in one case and green-dimmed OR lights in two cases due to patient photosensitivity after Gliolan administration; this primarily affected the researcher's ability to visualise landmark placement rather than HL2 performance. Interactions with anaesthesiologists occasionally interfered with landmark placement (n=3). For example, when tape obstructed the subnasale or the breathing tube needed repositioning, though system stability was generally maintained. Patient characteristics, such as small head size or loose skin, complicated registration in three cases. Poor CT quality (e.g., loose fiducials or flattened ears; n=3) and occasional incomplete automatic segmentation of the superior fiducial (n=2) also required attention. Finally, a few issues arose from wrongly planned landmarks, mostly due to user inexperience (n=2).

3.3 DISCUSSION

This chapter presented the first prospective evaluation of the Lumi AR-based registration workflow in a clinical setting on real patients using CT-based holograms and anatomical landmarks. Following a formative phase focused on technical refinement, the summative pilot study evaluated the accuracy and usability of anatomical landmark registration. Overall, the study demonstrated that CT-based anatomical landmark registration using the Lumi AR workflow is generally feasible within the OR and sufficiently accurate to support

and potentially improve EVD placement. Meanwhile, robustness remained constrained by hardware limitations: during the pilot study, three critical failures occurred, requiring restarts before the workflow could be completed. Such interruptions negatively affect usability and represent an important barrier to routine clinical deployment. These limitations are primarily attributable to constraints of the AR hardware rather than the registration methodology itself. In the pilot study (n=11), the system achieved a mean TRE of 4.9 mm (SD, 2.1 mm; median 4.8 (IQR, 2.0)) at the nasion and fiducial-based mean TREs ranging from 4.9 to 7.4 mm (median 4.8 to 6.9 mm). These values are just below and around the 5 mm accuracy hypothesised as necessary for safe EVD placement.

3.3.1 Interpretation of key findings

Several factors likely influenced these results. First, reliance on anatomical landmarks, particularly those on soft tissue, introduced a larger fiducial localisation error (FLE). FLE is defined as the Euclidean distance between a virtual and corresponding physical point, and FRE is a combination of the FLEs at all registration points (21). Unlike artificial markers, anatomical landmarks are subject to observer interpretation. Even for a single observer, consistently indicating the validation point, the nasion in this study, with the pointer, can still be challenging, as it is not a discrete landmark. Ideally, a dedicated fiducial marker could have reduced this ambiguity. However, this was not feasible because not all patients underwent CT-based neuronavigation. In the OR environment, additional practical factors further influence landmark accessibility: the nasion is generally unobstructed and easily visible, whereas auricular landmarks may be obscured by hair, and caution is needed for landmarks around the vulnerable eyes. Unfortunately, in this study, the head device sometimes partially covered the nasion, and as a result, accessibility varied with patient-specific anatomy. Furthermore, translating landmarks from the virtual model to the patient might have been challenging, as the model was prepared before seeing the patient in the OR. Minor differences between preoperative planning and the patient's actual anatomy could have made landmark placement more difficult. However, this reflects a typical workflow and is likely to be used in future procedures as well. Additionally, gloves reduce tactile feedback, and medical devices, such as endotracheal tubes or fixation tape, can interfere with facial landmarks.

It is also important to note that TRE was chosen as the primary metric because it provides the most clinically relevant measure available, but it is not a perfect surrogate. Accurate EVD placement depends not only on surface alignment but primarily on the offset of the drain tip within the ventricles and the trajectory's angulation. Moreover, because TRE was measured on the skin surface, any surface deviations are geometrically magnified relative to deeper intracranial points. This means that the surface TRE overestimated the potential error at the ventricular target.

Another contributing factor might have been the ergonomic and line-of-sight constraints associated with the fiducial positions. The mean and median TRE at fiducial1 were clearly higher than those at fiducial2. This difference may be related to the relative position of

fiducial1 with respect to the head-mounted device. This required the researcher to adopt an awkward, unstable head-and-body posture to simultaneously maintain optical tracking of both the marker on the pointer and the head-mounted device.

When considering the secondary outcomes, the key finding was that the system demonstrated stable "goodness of fit" comparable to preclinical phantom studies, high visual accuracy ratings (median, 5/5), and an efficient workflow (mean 4 min 36 s, SD 1 min 12 s; median 5 min 6 s, IQR 1 min 48 s). From a workflow perspective, registration times consistently remained well below 10 minutes. Given that AR guidance may facilitate more accurate drain placement and reduce the number of insertion attempts, the added setup time is unlikely to represent a major limitation. Nevertheless, procedural time remains a critical consideration, as EVD placement is often performed in emergency settings where no setup time is currently required. Regarding time, the only bottleneck identified is the additional delay associated with restarting the system after a software crash or extreme instability. The observed mean FRE of 4.0 mm (SD, 0.7 mm) is remarkably consistent with the 4.00 mm mean (SD, 1.16 mm) FRE observed in the prior phantom study (24). These results suggested that both user point localisation and the registration transformation calculation are non-inferior, even when transitioning from a rigid phantom setup to patients in the OR. A notable difference between the studies is that the phantom study did not include the subnasale landmark and used the tragus instead of the auricular roots. In theory, this could give the present study an advantage, as the auricular roots are more rigid and anatomically stable, and the subnasale improves the spatial distribution of registration points. However, FRE measures only the root mean square error of the points used in the registration itself and can be misleading: a low FRE may occur even if the overall registration is slightly off, as long as the errors are spread evenly across landmarks (21). Furthermore, there was a notable discrepancy between the TRE values and the visual ratings. Visual ratings were inherently subjective, difficult to compare across studies, and inconsistently reported in the literature; they were therefore included only as supportive information. In most cases, the hologram appeared well aligned with the patient's facial features, even when TRE values were both relatively low and relatively high. These differences are not entirely unexpected, since visual ratings and FRE capture global alignment, whereas TRE assesses accuracy at a single, specific location, making it more sensitive to localised misalignments.

Regarding system stability, a largely binary performance pattern was observed: the AR system either functioned reliably or exhibited obvious instability, including drift, menu displacement, or software crashes. Importantly, no instances of subtle failure were encountered in which the system appeared stable but produced inaccurate registration. From a clinical perspective, clear instability is preferable, as it prompts the user to abort or restart the procedure rather than proceed based on misleading guidance. Nevertheless, during the 11 measurements in this study, 3 critical failures occurred. This frequency remains too high to support routine clinical implementation. The observed "environment unstable" errors were likely related to the dynamic OR environment, underscoring the sensitivity of HL2 inside-out tracking to changes in lighting, motion, and surrounding personnel.

Although the workflow and anatomical landmark registration were generally feasible, the primary source of variability and occasional failure appeared to originate from the HL2 itself. Tracking performance varied between sessions, such that successful registration often depended on optimal device behaviour at the time of use. This hardware-dependent variability highlights that, even with careful landmark placement, overall registration accuracy remains constrained by the stability of the AR tracking system.

3.3.2 Comparison with existing literature

Literature on AR point-based registration using CT-derived models and anatomical landmarks remains scarce, particularly in real intraoperative settings. The small patient cohorts and the diversity of outcome reporting further restrict cross-comparisons. Most available evidence concerns fiducial marker–based registration or other navigation systems, such as optical tracking. Outside the AR domain, Woerdeman et al. reported, for instance, TRE values ranging from 4.03 to 6.03 mm for optical tracking using anatomical landmarks and CT scans (28). The results of this pilot study fall approximately within the same range of values. They also highlighted the superior accuracy of fiducial markers compared with anatomical landmarks in an optical navigation system (28). Unfortunately, fiducial marker registration is not feasible in emergency settings, such as EVD placement.

Two recent unpublished studies evaluated the use of these fiducial markers for AR registration in cohorts of 37 patients, using the HL2 and Lumi software. The MRI-based study reported a mean FRE of 4.6 mm (SD, 1.4 mm) and a mean TRE of 5.6 mm (SD, 3.0 mm), while the CT-based study demonstrated a lower mean FRE of 3.0 mm (SD, 1.3 mm) with a comparable mean TRE of 6.1 mm (SD, 3.1 mm) (31, 32). This difference aligns with the higher spatial resolution and lower geometric distortion of CT scans, which enable more reliable annotation of surface anatomical landmarks than MRI. This is particularly the case for landmarks closely related to underlying bony structures. In the present pilot study, the mean FRE of 4.0 mm (SD 0.7 mm) fell between those reported in earlier Lumi studies. This was expected, as CT allows more accurate visualisation. However, subjective identification of anatomical landmarks introduces variability and reduces accuracy compared with fiducial marker–based registration. Notably, the TRE at the nasion (mean 4.9 mm, SD 2.1 mm) had a lower mean and SD than those reported in fiducial-based studies using Lumi. This may be explained by the fact that the nasion lay along the same midline axis as the registration points, resulting in a favourable geometric configuration. TRE was also assessed at two fiducial locations; however, the limited number of measurements (n=3 and n=4) precluded meaningful comparison. The higher TRE observed at these targets likely reflects their position outside the axis plane of the registration landmarks.

3.3.3 Limitations

Several limitations restrict the generalizability of these results. The most significant limitations were the small sample size (n=11) and the lack of enough reliable validation points for TRE calculation. Validation primarily relied on surface landmarks, which are

susceptible to soft-tissue variability, observer-dependent interpretation, and, in this study, occasional occlusion by the head-mounted display. Moreover, surface-based TRE reflects registration accuracy only at the skin level. In contrast, the clinically relevant outcome for EVD placement is the accuracy of the trajectory and drain tip within the intracranial target.

A third limitation is that the non-inferiority analysis compared the OR results with those from the phantom study using only FRE. While this confirmed that the registration algorithm performs consistently, relying solely on FRE does not guarantee that clinical outcomes (drain placement accuracy) would be non-inferior. TRE, along with visual accuracy ratings, remains critical for a comprehensive assessment. Ideally, the phantom study used for non-inferiority analysis should have included TRE as an additional outcome measure.

3.3.4 Future Directions

To establish a definitive conclusion regarding the accuracy of anatomical landmark registration, future research must include a fully powered cohort. Subsequent studies should prioritise patients undergoing standard neuronavigation; the inclusion of artificial fiducials near Kocher's point provides better validation points for calculating TRE at the exact site of surgical entry, rather than relying on distant surface landmarks.

Looking forward, the instability of soft-tissue landmarks remains a bottleneck. Investigating alternative registration strategies, such as markerless surface matching or automated detection of anatomical landmarks, could reduce interobserver variability. Finally, addressing the system instability is also a prerequisite for further use in clinical, high-stakes environments. At the moment, the observed instability of the HL2 prevents reliable and trustworthy use of this workflow for EVD placement.

3.4 CONCLUSION

The Lumi AR workflow demonstrated that anatomical landmark registration with CT-based holograms is feasible and can provide rapid, intuitive visualisation for EVD placement. This offers a promising alternative to the current freehand standard of care. While registration accuracy requires further validation, the approach itself is sound. The main limitations observed were hardware-related: instability and session-to-session variability of the HL2 constrained system reliability. These findings indicate that the concept of AR-assisted EVD placement using anatomical landmarks is viable, but safe and consistent clinical use will require more robust AR hardware. Overall, this study supports further exploration of landmark-based AR guidance independent of current device limitations.

IV. Discussion & Conclusion

This master's thesis evaluated whether CT-based anatomical landmark registration using the Lumi AR workflow is sufficiently accurate, robust, and feasible to support EVD placement, thereby improving drain placement accuracy. By addressing both AI-assisted landmark annotation and the performance of the complete AR-based registration workflow in the OR, this work provided an integrated assessment of the technical and clinical readiness of this approach.

4.1 AI-ASSISTED ANATOMICAL LANDMARK ANNOTATION

The first subgoal was to assess the clinical acceptability of AI-generated anatomical landmarks. The results demonstrated that, while AI-based landmarking can provide a useful initial estimate, it is not yet sufficiently reliable for direct clinical use without manual correction. Landmark acceptability varied systematically across anatomical regions: rigid, well-defined bony landmarks, such as the nasion, showed higher interobserver agreement and a lower adjustment rate, whereas soft-tissue landmarks, such as those around the eyes, showed greater variability. These findings emphasised that landmark reliability is influenced not only by algorithm performance but also by the ambiguity of surface anatomy and the quality of the skin segmentation. As such, AI-assisted landmarking is not currently suitable for direct use in time-critical clinical settings, such as EVD placement.

4.2 REGISTRATION ACCURACY IN THE OR

The second subgoal concerned the accuracy of anatomical landmark–based registration in a real clinical environment. This accuracy can be understood as a causal chain: the more accurate the annotated landmarks, the more accurate the registration and resulting visualisation, which in turn increases the likelihood of achieving a favourable Kakarla Grade I (optimal drain placement) rate. Although the pilot study was not powered for definitive accuracy claims, the observed TRE values suggested that point-based registration using CT-derived anatomical landmarks can achieve accuracy that is likely clinically acceptable for EVD placement. Given the relatively large target volume of the ventricular system, the measured TREs fall within a range that could support safe drain placement. Importantly, these results demonstrated that the registration methodology itself is technically sound when translated from a phantom setup to real patients in the OR. While this AR-guided workflow did not achieve the 1–2 millimetre precision of stereotactic systems, it is expected to provide practical accuracy comparable to other traditional guidance methods, such as mechanical guides and ultrasound. Besides, it has the added advantage of direct 3D visualisation of the patient's anatomy within the operative field.

4.3 WORKFLOW FEASIBILITY

With respect to workflow feasibility, AR-based registration was consistently completed within a short timeframe, typically in five minutes. This aligns well with the time constraints of acute EVD placement, even in emergency settings. Integration with the hospital PACS, the absence of an external navigation screen, and the ability to perform registration without rigid head fixation further support the workflow's clinical practicality. Due to its speed and minimal equipment requirements, the workflow may also be applicable outside the OR, for example, in the ICU. As point-based registration is already familiar to most neurosurgeons, the primary novelty lies in its execution through an AR head-mounted display rather than in the registration process itself.

4.4 SYSTEM ROBUSTNESS

The fourth subgoal addressed system robustness, which proved to be the most significant limitation. During the pilot study, several system instabilities were observed, ranging from transient issues such as hologram dropouts and menu displacement to more severe events, including full software crashes. Although most events were recoverable and allowed measurements to be completed, these interruptions nonetheless negatively affected usability. Importantly, all failures were overt rather than subtle: no instances were observed in which the system appeared stable while producing inaccurate registration. From a clinical safety perspective, such explicit failures are preferable to silent inaccuracies, as they prompt the user to restart or abort the procedure. Notably, these limitations appeared primarily attributable to the AR hardware and tracking stability rather than to the registration methodology itself. Consequently, they should not be interpreted as a failure of the AR-guided concept, but rather as a constraint of the current generation of hardware. Overall, the observed frequency of (critical) failures remains too high to support clinical implementation using the HL2.

4.5 FUTURE DIRECTIONS

Based on the earlier mentioned IDEAL framework, the AR-guided EVD workflow evaluated in this thesis remains within Stage 2a (Development). While the underlying registration methodology was conceptually sound and demonstrated clinically acceptable accuracy and feasibility, progression to Stage 2b (Exploration) was limited by insufficient system robustness. Since Stage 2b requires stable technology to enable meaningful evaluation in larger cohorts, further technological improvements are necessary before such studies can be conducted (25). Accordingly, the evolution of the Lumi AR workflow should thus proceed in two phases: technological hardening and clinical validation.

First, development must prioritise hardware transition. The observed instabilities indicate that the HL2 is insufficient for (high-stakes) surgical environments. Future iterations should migrate the workflow to surgical-specific hardware or more robust headsets with superior tracking stability. Concurrently, landmark annotation should be fully integrated into the web interface. While an updated AI tool could eventually be incorporated, this is not urgent,

as annotation in the web interface already significantly improves usability compared with annotation in the HL2 application. Additionally, designing a more ergonomic head-mounted reference device and pointer will be important to enhance usability and patient comfort further. Standardised CT acquisition protocols should also be explored to ensure optimal coverage of surface landmarks and imaging quality. Complete standardisation may be challenging across clinical settings due to scanner availability, patient conditions, or emergency settings. However, establishing minimum imaging requirements, such as slice thickness, head orientation, and FOV, could substantially improve consistency and registration reliability.

Although the current findings motivate a transition to more robust AR hardware, expanding the pilot dataset using the HL2 may still be valuable to further characterise expected accuracy ranges and methodological performance of CT-based anatomical landmark registration. However, definitive conclusions regarding system robustness and clinical readiness should be reserved for studies conducted on new hardware. Accuracy validation should, where possible, incorporate fiducial markers to strengthen the reliability of TRE measurements and to improve comparability with established MRI/CT-based registration approaches using fiducials (31, 32).

Subsequently, a prospective cohort study of AR-guided EVD placement using new hardware should be conducted. This design would allow evaluation of the workflow under clinical conditions, without exposing patients to potentially inferior freehand placement, especially given the preliminary evidence that guided techniques improve accuracy (14, 24). To strengthen methodological rigour, outcomes could be compared with a matched cohort of historical freehand EVD placements derived from hospital records. Outcome measures should extend beyond registration accuracy, quantified by TRE, to include clinically meaningful endpoints, such as drain-tip deviation from the intended ventricular target and angular deviation. This can be calculated by comparing postoperative CT scans with preoperative planning data. Ultimately, demonstrating that AR guidance consistently achieves a high proportion of Kakarla Grade 1 placements with fewer insertion attempts will be essential for establishing this technology as a new standard of care for EVD placement.

4.6 CONCLUSION

In conclusion, AR-guided EVD placement using CT-based anatomical landmark registration is a promising approach with clear clinical potential. Although AI-assisted landmark annotation is not yet ready for use and current AR hardware limits system robustness, the underlying registration workflow proved accurate and feasible. These findings support continued development and validation of AR-guided EVD placement using more stable hardware to improve robustness while preserving the accuracy and workflow feasibility already demonstrated in this thesis.

References

1. Willems PWA, Van Der Sprenkel J, Tulleken CAF, Viergever M, Taphoorn MJB. Neuronavigation and surgery of intracerebral tumours. J Neurol. 2006;253(9):1123-36.

2. Wagner W, Tschiltschke W, Niendorf WR, Schroeder HW, Gaab MR. Infrared-Based Neuronavigation and Cortical Motor Stimulation in the Management of Central-Region Tumors. Stereotact Funct Neurosurg. 1997;68(1-4 Pt1):112-6.

3. Mongen MA, Willems PWA. Current accuracy of surface matching compared to adhesive markers in patient-to-image registration. Acta Neurochir (Wien). 2019;161(5):865-70.

4. Sekula RF, Cohen DB, Patek PM, Jannetta PJ, Oh MY. Epidemiology of ventriculostomy in the United States from 1997 to 2001. Br J Neurosurg. 2008;22(2):213-8.

5. O'Neill BR, Velez DA, Braxton EE, Whiting D, Oh MY. A Survey of Ventriculostomy and Intracranial Pressure Monitor Placement Practices. Surg Neurol. 2008;70(3):268-73.

6. Kakarla UK, Kim LJ, Chang SW, Theodore N, Spetzler RF. Safety and accuracy of bedside external ventricular drain placement. Neurosurgery. 2008;63(1 Suppl 1):ONS162-6.

7. Connolly ES, McKhann II GM, Komotar RJ, Mocco J, Choudhri AF. Fundamentals of Operative Techniques in Neurosurgery. Thieme; 2010.

8. Thamjamrassri T, Yuwapattanawong K, Chanthima P, Vavilala MS, Lele AV, Collaborators ES. A Narrative Review of the Published Literature, Hospital Practices, and Policies Related to External Ventricular Drains in the United States: The External Ventricular Drain Publications, Practices, and Policies (EVDPoP) Study. J Neurosurg Anesthesiol. 2022;34(1):21-8.

9. Dawod G, Henkel N, Karim N, Caras A, Qaqish H, Mugge L, et al. Does the Setting of External Ventricular Drain Placement Affect Morbidity? A Systematic Literature Review Comparing Intensive Care Unit versus Operating Room Procedures. World Neurosurg. 2020;140:131-41.

10. Fisher B, Soon WC, Ong J, Chan T, Chowdhury Y, Hodson J, et al. Is Image Guidance Essential for External Ventricular Drain Insertion? World Neurosurg. 2021;156:e329-e37.

11. Augmedit. State of the Art Analysis: Freehand Placement of EVDs. Part of CE marking process; unpublished, internal document. 2025.

12. Nawabi NLA, Stopa BM, Lassaren P, Bain PA, Mekary RA, Gormley WB. External ventricular drains and risk of freehand placement: A systematic review and meta-analysis. Clin Neurol Neurosurg. 2023;231:107852.

13. Stuart MJ, Antony J, Withers TK, Ng W. Systematic Review and Meta-analysis of External Ventricular Drain Placement Accuracy and Narrative Review of Guidance Devices. J Clin Neurosci. 2021;94:140-51.

14. Goossens MC. Real-time Guidance for Ventricular Puncture: A Scoping Review of Available Techniques and Placement Accuracy [unpublished manuscript]. Educational program Technical Medicine; Leiden University Medical Center, Delft University of Technology & Erasmus University Medical Center Rotterdam, The Netherlands. 2025.

15. Fick T, Van Doormaal JAM, Hoving EW, Regli L, Van Doormaal TPC. Holographic patient tracking after bed movement for augmented reality neuronavigation using a head-mounted display. Acta Neurochir (Wien). 2021;163(4):879-84.

16. Van Doormaal TPC, Van Doormaal JAM, Mensink T. Clinical Accuracy of Holographic Navigation Using Point-Based Registration on Augmented-Reality Glasses. Oper Neurosurg. 2019;17(6):588-93.

17. Van Doormaal JAM, Fick T, Ali M, Köllen M, Van Der Kuijp V, Van Doormaal TPC. Fully Automatic Adaptive Meshing Based Segmentation of the Ventricular System for Augmented Reality Visualization and Navigation. World Neurosurg. 2021;156:e9-e24.

18. Fick T, Van Doormaal JAM, Tosic L, Van Zoest RJ, Meulstee JW, Hoving EW, et al. Fully automatic brain tumor segmentation for 3D evaluation in augmented reality. Neurosurg Focus. 2021;51(2):E14.

19. van Doormaal J, van Doormaal T. Augmented Reality in Neurosurgery. In: Di Ieva A, Suero Molina E, Liu S, Russo C, editors. Computational Neurosurgery. Advances in Experimental Medicine and Biology. 1462: Springer; 2024. p. 351-74.

20. De Boer M, Van Doormaal JAM, Köllen MH, Bartels LW, Robe PAJT, Van Doormaal TPC. Fully automatic anatomical landmark localization and trajectory planning for navigated external ventricular drain placement. Neurosurg Focus. 2025;59(1):E14.

21. Fitzpatrick JM, editor Fiducial registration error and target registration error are uncorrelated. Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling; 2009: SPIE

22. Campisi B, Costanzo R, Gulino V, Avallone C, Noto M, Bonosi L, et al. The Role of Augmented Reality Neuronavigation in Transsphenoidal Surgery: A Systematic Review. Brain Sciences. 2023;13(12):1695.

23. Kim S, Kim K. How to use neuronavigation for the brain. Neurofunction. 2021;17(2):126–32.

24. van Doormaal JAM, Fick T, Meulstee JW, Kos TM, Bot M, O'Donnell P, et al. External Ventricular Drain Placement Using Active Augmented Reality Guidance: A Proof of Concept of a Clinically Integrable System. Operative Neurosurgery. 2025;00:1-8.

25. McCulloch P, Altman D, Campbell W, Flum D, Glasziou P, Marshall J, et al. No surgical innovation without evaluation: the IDEAL recommendations. The Lancet. 2009;374(9695):1105–12.

26. Kendrick A. Formative vs. Summative Evaluations: Nielsen Norman Group (NN/g); 2019 [cited 2026 2 Jan]. Available from: https://www.nngroup.com/articles/formative-vs-summative-evaluations/.

27. Microsoft Corporation. HoloLens 2 Industrial Edition – specificaties en functies[cited 2025 30 Dec]. Available from: https://www.microsoft.com/nl-NL/p/hololens-2-industrial-edition/8mqn5pzp01x5?activetab=pivot:overzichttab.

28. Woerdeman PA, Willems PWA, Noordmans HJ, Tulleken CAF, Berkelbach van der Sprenkel JW. Application accuracy in frameless image-guided neurosurgery: a comparison study of three patient-to-image registration methods. J Neurosurg. 2007;106:1012–6.

29. Shi Z, Peng Y, Gao X, Chen S, Chen G, Pan G, et al. Translating high-precision mixed reality navigation from lab to operating room: design and clinical evaluation. BMC Surgery. 2025;25:Article Number 331.

30. Qi Z, Li Y, Xu X, Zhang J, Li F, Gan Z, et al. Holographic mixed-reality neuronavigation with a head-mounted device: technical feasibility and clinical application. Neurosurgical Focus. 2021;51(2):E22.

31. Kos TM, Maathuis WD, Willems P, Bartels LW, Robe PA, van Doormaal TPC. Comparison of a Novel Augmented Reality–Based and a State-of-the-Art Infrared Optical–Based Image-to-Patient Registration Method in Neurosurgery [Unpublished]. 2025.

32. O'Donnell PT, Goossens MC, Razoux Schultz AT, Dreissen YEM, Schuurman PR, Bot M. Augmented Reality-based registration using fiducial markers and CT imaging [Unpublished]. 2025.

Appendices

A. LUMI SOFTWARE BUILDS AND CHANGE LOG

Overview

Situation	Version	Notes
Chapter 2	v.1.0.19859.0	Added the AI landmarking feature
Chapter 3; formative phase	v.1.0.20586.0	Initial formative evaluation build
Chapter 3; summative phase	v.1.0.20852.0	Updated based on insights from the formative phase
	v.1.0.20950.0	This build was released to fix a bug identified in v.1.0.20852.0 during this phase. As the bug did not affect the EVD workflow or registration process, the updated version was used to avoid working with a known suboptimal build.

Key Updates v.1.0.20586.0 (formative) → v.1.0.20852.0 (summative)

- **Frame rate alerts:** Low frame rates trigger a real-time warning and temporarily halt registration until the system is stable again to ensure tracking accuracy.
- **Spatial awareness alerts:** Users receive instructions to look around the room to restore hologram stability if spatial awareness is lost.
- **User detection warnings:** Improved warning notifications when the visor is flipped up or the user is not recognised.
- **EVD trajectory adjustment & recalibration:** Users can adjust the virtual trajectory and recalibrate the reference marker again later in the workflow.
- **Bug fixes:** Minor issues resolved, including preventing retention of previous calibration values and correct display of calibration values.
- **Enhanced logging:** More detailed tracking of user actions to aid troubleshooting and development; automatic upload to dashboard or crash logs sent to Augmedit.
- **Target point locking:** The target point is locked in later workflow steps to prevent accidental movement.
- **UI improvements:** Text and graphics updated for clarity; 'Edit Mode' no longer auto-starts; visual confirmation when sufficient landmarks are placed.

B. PARTICIPANT INSTRUCTIONS

Experiment 1: Clinical acceptability of AI-generated Anatomical Landmarks

In this experiment, you will review anonymised CT-based 3D patient models containing automatically placed anatomical landmarks. These landmarks are visualised in Lumi on the HoloLens 2 (HL2). For each landmark, you will decide whether its placement is acceptable for clinical use or whether adjustment would be required. The purpose of this study is to assess how clinicians perceive the accuracy of AI-placed landmarks on 3D patient models.

Time per session
~25–35 minutes total.

Materials

- HL2 with the Lumi application.
- 15 anonymised CTs with skin segmentations and AI-placed landmarks.

Background
Seven landmarks are generated automatically by the AI algorithm (see Figure 1):

Point	Anatomical location
1	Nasion
2	Medial canthus left
3	Medial canthus right
4	Lateral canthus left
5	Lateral canthus right
6	Auricular root left
7	Auricular root right

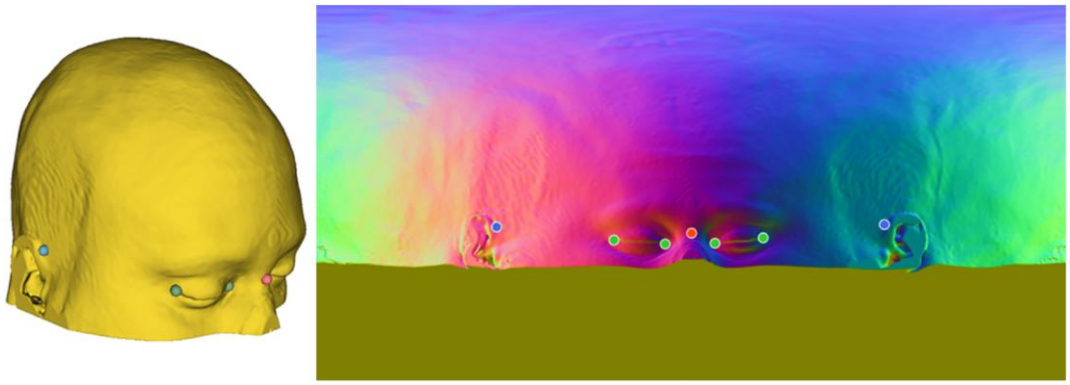

Figure 1: Visualization of the seven anatomical landmarks (1)

In clinical practice, these reference points are used for point-based registration. Point-based registration refers to the process of aligning a patient's anatomy in real life with their virtual

imaging data (e.g., CT or MRI) by selecting specific anatomical points in both worlds. This is commonly done in the operating room (OR) using systems such as Brainlab or StealthStation. For this study, you are asked to judge whether each AI-placed landmark on the virtual skin model is acceptable as-is, or if you would adjust it before continuing with registration, based on your own experience.

Some criteria to keep in mind when judging a landmark:

- Must correspond to one of the seven listed anatomical locations.
- Should lie on a clearly identifiable surface that could be located on the real patient, too.

Participant Instructions

1. **Onboarding:**
    - You will receive a short briefing.
    - Fit the HL2 comfortably, adjust it to your eyes and confirm that the display text is clear in the centre of view.
2. **Main evaluation:**
   For each patient case:
    a. Load the case provided by the researcher.
    b. Explore the 3D model freely (walk around, zoom, rotate as needed). Make sure the landmarks and corresponding labels are switched on!
    c. For each landmark, decide whether it is:
        - **Accept** = placement is sufficient for registration; no adjustment needed.
        - **Adjust** = placement is not sufficient; you would reposition before registration.

       Simply say out loud to the researcher which landmark numbers you would adjust.
    d. Proceed until all landmarks in the case are classified.
    e. Confirm completion; continue to the next case.

References

A. de Boer M, van Doormaal JAM, Köllen MH, Bartels LW, Robe PAJT, van Doormaal TPC. Fully automatic anatomical landmark localization and trajectory planning for navigated external ventricular drain placement. Neurosurg Focus. 2025 Jul;59(1):E14.

C. INTEROBSERVER AGREEMENT PER HOLOGRAM AND ANATOMICAL LOCATION

ID	Partial agreement (%) (Median (IQR)	Full agreement (%) (Mean [95% CI])
Hologram*		
1	75(25.0)	42.9 [15.8–75.0]
2	75(25.0)	42.9 [15.8–75.0]
3	100(25.0)	57.1 [25.0–84.2]
4	100(25.0)	71.4 [35.9–91.8]
5	100(0.0)	85.7 [48.7–97.4]
6	100(0.0)	85.7 [48.7–97.4]
7	100(12.5)	71.4 [35.9–91.8]
8	75(12.5)	28.6 [8.2–64.1]
9	75(12.5)	28.6 [8.2–64.1]
10	100(25.0)	57.1 [25.0–84.2]
11	100(0.0)	85.7 [48.7–97.4]
12	75(25.0)	28.6 [8.2–64.1]
13	100(0.0)	100 [64.6–100]
14	100(12.5)	71.4 [35.9–91.8]
15	100(25.0)	57.1 [25.0–84.2]
Anatomical location†		
1 – nasion	100(0.0)	93.3 [70.2–98.8]
2 – medial canthus (l)	100(0.0)	80.0 [54.8–92.9]
3 – medial canthus (r)	75(25.0)	46.7 [24.8–69.9]
4 – lateral canthus (l)	75(25.0)	46.7 [24.8–69.9]
5 – lateral canthus (r)	100(25.0)	60.0 [35.7–80.2]
6 – auricular root (l)	75(25.0)	46.7 [24.8–69.9]
7 – auricular root (r)	100(25.0)	53.3 [30.1–75.2]

Abbreviations: *l = left; r = right; IQR = interquartile range; CI = confidence interval.*

** Each hologram included 7 landmarks, so agreement was assessed across 7 elements per hologram.*

† Each anatomical location was evaluated across 15 different holograms, so agreement reflects assessments of 15 instances per location.

D. NON-INFERIORITY ANALYSIS AND SAMPLE SIZE CALCULATION

Non-inferiority margin

The phantom study reported a mean FRE:

$$\bar{\mu}_{phantom} = 4.00\ mm$$

To define a clinically acceptable margin (Δ), 20% of the phantom mean was chosen:

$$\Delta = 0.20 \cdot \bar{\mu}_{phantom} = 0.80\ mm$$

Hypotheses

The non-inferiority test compares the true difference in means $(\bar{\mu}_{OR} - \bar{\mu}_{phantom})$ to this margin:

- **Null Hypothesis (H_0):** OR registration is inferior. The true mean difference is greater than or equal to the margin.

$$H_0: \bar{\mu}_{OR} - \bar{\mu}_{phantom} \geq \Delta$$

- **Alternative Hypothesis (H_1):** OR registration is non-inferior. The true mean difference is less than the margin.

$$H_1: \bar{\mu}_{OR} - \bar{\mu}_{phantom} < \Delta$$

The OR registration is considered non-inferior if the upper bound of the two-sided 90% CI for the mean difference $(\bar{\mu}_{OR} - \bar{\mu}_{phantom})$ is below 0.80 mm.

Sample size calculation

Rationale

The sample size was calculated using a two-sample non-inferiority Z-test. While the final analysis utilises a t-distribution, the Z-approximation is the standard convention for planning and provides a transparent estimation of the required enrolment.

To ensure a conservative estimate, the sampling uncertainty from the completed phantom study (n=20) was incorporated into the total variance. Because this benchmark is derived from a limited sample, the sampling variance of its mean must be accounted for; ignoring this uncertainty would lead to an underestimation of the required sample size for the clinical study.

The standard Z-statistic for comparing two means is:

$$z = \frac{\left(\bar{\mu}_{OR} - \bar{\mu}_{phantom}\right) - \left(\mu_{OR} - \mu_{phantom}\right)}{\sqrt{\frac{\sigma_{OR}^2}{n_{OR}} + \frac{\sigma_{phantom}^2}{n_{phantom}}}},$$

where:

- $\bar{\mu}_{\text{OR}}$ is the observed mean FRE in the OR study.
- $\bar{\mu}_{\text{phantom}}$ is the mean FRE reported in the phantom study.
- $\mu_{\text{OR}} - \mu_{\text{phantom}}$ is the expected true difference in means (for non-inferiority, usually o).
- σ_{OR}^2 is the population variance of the OR measurements (or best available estimate).
- $\sigma_{\text{phantom}}^2$ is the population variance of the phantom study measurements.
- n_{OR} is the planned sample size for the OR study.
- n_{phantom} is the sample size of the phantom study.

For sample-size planning, the denominator determines the required precision of the estimate. It is also known that the variance of any sample mean is calculated as the population variance divided by the sample size:

$$\text{V} = \frac{\sigma^2}{n}$$

Assumptions:

- Non-inferiority margin:

$$\Delta = 0.80\ mm$$

- Expected true difference:

$$\bar{\mu}_{OR} - \bar{\mu}_{phantom} = 0\ mm$$

- Standard deviation (estimated from phantom study):

$$\sigma_{OR} = \sigma_{phantom} = s_{phantom} = 1.16\ mm$$

- Significance level:

$$\alpha = 0.05, Z_{1-\alpha} = 1.645$$

- Power:

$$1 - \beta = 0.80, Z_{1-\beta} = 0.84$$

Step 1 – Variance contribution of the phantom study

$$V_{phantom} = \frac{\sigma^2_{phantom}}{n_{phantom}} = \frac{1.16^2}{20} = 0.0673$$

Step 2 – Required total variance for non-inferiority

The required variance of the difference in means follows from the Z-test expression:

$$V_{required\ diff} = \frac{(\Delta - (\mu_{OR} - \mu_{phantom}))^2}{(Z_{1-\alpha} + Z_{1-\beta})^2} = \frac{(0.80 - 0)^2}{2.485^2} = 0.1036$$

This represents the maximum allowable variance of the difference between OR and phantom means while retaining 80% power.

Step 3 – Allowable variance contribution from OR study

$$V^*_{OR} = V_{required\ diff} - V_{phantom} = 0.1036 - 0.0673 = 0.0363$$

Step 4 – Solve for the required OR sample size

$$n_{OR} = \frac{\sigma^2_{OR}}{V^*_{OR}} = \frac{1.16^2}{0.0363} = 37.0 = 37\ patients$$

Pilot study considerations

Because the feasible enrolment for this thesis was limited to 10–15 patients, the present study is underpowered for a full non-inferiority conclusion. All CIs and hypothesis evaluations should therefore be interpreted as exploratory, and additional participants are required to complete the planned analysis.

Confidence interval construction

Rationale

While sample size planning relied on the normal (Z) approximation, the analysis employs the t-distribution to account for the small pilot sample. The Welch Two-Sample t-test is used to accommodate unequal variances between the groups; the static phantom measurements (n=20) are expected to have a different spread than the patient data (n=10–15). The Satterthwaite approximation is used to calculate the effective degrees of freedom (df) in this unequal-variance, unequal-sample-size scenario.

Calculation

To assess non-inferiority, the upper bound of a two-sided 90% CI for the difference in means $(\mu_{\text{OR}} - \mu_{\text{phantom}})$ was calculated using the Welch framework:

$$CI\ (upper\ bound) = (\bar{\mu}_{OR} - \bar{\mu}_{phantom}) + t^* \cdot \sqrt{\frac{s^2_{OR}}{n_{OR}} + \frac{s^2_{phantom}}{n_{phantom}}},$$

where:

- $\bar{\mu}_{OR} - \bar{\mu}_{phantom}$ is the observed difference between the two sample means,
- t^* is the critical value from the two-sided t-distribution corresponding to a 90% confidence level and the effective degrees of freedom,
- s^2_{OR} and s^2_{phantom} are the sample variances,
- n_{OR} and n_{phantom} are the respective sample sizes.

The effective degrees of freedom for the Welch t-test were calculated using the Satterthwaite approximation:

$$df_{\mathrm{Satt}} = \frac{\left(\frac{s^2_{\mathrm{OR}}}{n_{\mathrm{OR}}} + \frac{s^2_{\mathrm{phantom}}}{n_{\mathrm{phantom}}}\right)^2}{\frac{(s^2_{\mathrm{OR}}/n_{\mathrm{OR}})^2}{n_{\mathrm{OR}} - 1} + \frac{(s^2_{\mathrm{phantom}}/n_{\mathrm{phantom}})^2}{n_{\mathrm{phantom}} - 1}}$$

The critical t value was determined from the t-distribution using the calculated effective degrees of freedom (df) at the significance level $\alpha = 0.05$. The upper bound of this CI is then compared to the non-inferiority margin; if it is below $\Delta = 0.80$ mm, OR registration is considered non-inferior.

E. DETAILED RESULTS OF THE OR PILOT STUDY

FRE & TRE metrics

Patient	FRE (mm)	TRE-nasion (mm)	TRE-fiducial1 (mm)	TRE-fiducial2 (mm)
FORMATIVE PHASE				
1	5.0	5.2	-	-
2	3.9	NA	NA	NA
3	4.2	1.9	-	-
4	5.5	NA	-	-
5	4.9	5.3	5.7	6.9
SUMMATIVE PHASE				
1	3.2	4.6	6.1	2.9
2	3.6	0.5	NA	7.2
3	3.6	4.6	-	-
4	2.9	3.8	-	-
5	3.4	NA	-	-
6	4.9	3.6	-	-
7	3.6	5.1	-	-
8	4.4	7.1	-	-
9	5.0	8.0	9.4	3.8
10	4.1	6.1	6.9	5.8
11	5.0	5.7	-	-

Abbreviations: *FRE = fiducial registration error; mm = millimetres; TRE = target registration error; NA = not available (data expected but not obtained due to workflow interruption or technical issues).*

Non-inferiority analysis (FRE)

Mean OR (SD) (n=11)	Mean Phantom (SD) (n=20)	Difference in mean	df *	t_critical†	Upper bound 90% CI
3.98 (0.74)	4.00 (1.16)	-0.02	28.21	1.70	0.56

Abbreviations: *FRE = fiducial registration error; OR = operating room; SD = standard deviation; n = number; df = degrees of freedom; CI = confidence interval.*

** Degrees of freedom calculated using the Satterthwaite approximation for unequal variances.*

† Critical t-value for the upper bound of the two-sided 90% confidence interval.

Visual registration accuracy rating

Patient	Nose	Ear (l)	Ear (r)	Back of the head
FORMATIVE PHASE				
1	3	4	3	5
2	5	5	4	5
3	3	3	2	4
4	5	5	5	5
5	5	5	5	5
SUMMATIVE PHASE				
1	5	5	5	5
2	5	5	5	5
3	4	4	4	4
4	5	5	5	5
5	5	5	5	5
6	4	3	3	4
7	5	5	5	5
8	4	5	5	4
9	5	5	5	5
10	3	3	4	4
11	5	4	4	5

Abbreviations: *l = left; r = right.*
**Scored on a 5 point Likert-scale (1 = very poor, 5 = excellent).*

Registration time

Patient	Duration per workflow step (s)					
	Plan EVD trajectory	Plan landmarks	Prepare patient	Patient registration	Validate registration	Total
FORMATIVE PHASE						
1	7	5	14	163	46	**235**
2	9	6	30	305	-	**350**
3	16	9	49	345	21	**440**
4	17	4	19	158	22	**220**
5*	NA	NA	NA	NA	NA	**300**
SUMMATIVE PHASE						
1	46	15	51	147	59	**318**
2	21	5	39	130	45	**240**
3	14	9	29	231	20	**303**
4	14	7	45	216	26	**308**
5	12	6	33	344	19	**414**
6	11	16	23	121	14	**185**
7	6	2	73	86	33	**200**
8	11	8	29	208	48	**304**
9	14	7	45	173	68	**307**
10	15	5	43	221	18	**302**
11	NA	NA	18	116	17	**151**

Abbreviations: *s = seconds; EVD = external ventricular drain; NA = not available (data expected but not obtained due to workflow interruption or technical issues).*
** For this measurement, the software was unable to record individual step durations; only the total workflow time was captured manually.*