ADVT Assessment

Uses of the Microsoft Hololens and Augmented Reality in Neurosurgery Since 2018

Chris Cowl
Y3908147

Keywords

Microsoft Hololens; neurosurgery; neuronavigation; image-guided surgery; augmented-reality; mixed-reality; surgical accuracy.

Abstract

Accuracy in neurosurgery is paramount. Referring to 2D screens adjacent to the surgical field costs surgeons time and cognitive effort which could increase the chance of error. Augmented-reality (AR) Microsoft Hololens-based neuronavigation systems hope to aid surgeons by overlaying postoperative and intraoperative information within the surgical field, most notably by providing a 3D hologram of the anatomy with a surgical path to it, hopefully increasing the chance of a successful surgical outcome. N = 16, 80% of papers in this literature review show that accuracy is of vital interest, of which N = 13, 65% focus on accuracy of navigation and N = 3, 15% focus on accuracy of registration. N = 4, 20% focus on user experience of the AR system. Only N = 6, 30% of papers find the AR system is clinically accurate. More research is needed on real patients to test accuracy, with more ecological valid studies and higher participant numbers.

1 INTRODUCTION

The CT scanner, brought to hospitals in 1973, enables surgeons to x-ray patients and plan complex surgery with the use of many x-ray 'slices' of the human anatomy (Shaikhouni & Elder 2012). This early computerised navigation aided medical surgeons in performing surgery. As processing software and computing power have evolved and increased, image-guided surgery has now become common in many areas of neurosurgery and is particularly useful when operating deep in critical areas of the brain and spinal cord where mistakes and unnecessary damage caused by the surgeon can be costly (Grimson et al., 1999). With that being said, accuracy and therefore reduction of errors is clearly of utmost importance to surgeons with general medical errors in the US costing $17.1b in 2008 and around 22,000-29,000 Americans dying each year from preventable medical errors (Rodwin et al., 2020; Van Den Bos et al., 2011). Therefore it's clear that increasing accuracy and safety and reducing preventable medical error will continue to be important factors when a new technology is introduced to the operating theatre.

Surgeons have to think about what important structures surround the proposed preoperative navigation path so as to ensure that as little collateral damage is done and minimally invasive surgery (MIS) is achieved (Teo 2010). One way of aiding surgeons through the patient anatomy is by the use of 2D screens displaying preoperative MRI/CT scans adjacent to the surgical field. Another way is the use of an intraoperative o-arm 3D fluoroscopy technique to update the surgeon whilst operating (Thakkar et al., 2017). Both these techniques however, create a few challenges for the surgeon. Firstly, the surgeon has to continually turn their head to the 2D screen during navigation, they have to understand the surgical path ahead, turn their head back to the surgical field and refocus on the context of the anatomy. This creates a lot of cognitive cost and attention switching, and therefore presents a divided attention problem that can hinder surgeons (Ghazanfar et al., 2015). Research shows that having too many tasks on the go at any one time can force people to reach a maximum capability (Carswell et al., 2005) so it's a good idea to try to smooth surgeon workflow to reduce this. Secondly, the surgeon has to interpret a 2D image and then translate it to the context of the surgical field, and so there is potential for human error due to interruptions (Göras et al., 2019). Thirdly, in the case of the 3D fluoroscopy technique, the patient is subjected to doses of radiation, although lower than the previous 2D fluoroscopy technique (Thakkar et al., 2017), which is used intraoperatively to help the surgeon perform the procedure more accurately. Finally, the size of the equipment used to show this preoperative information can be large and from a practical and expense perspective, is systemically again something that can be improved upon. Therefore image-guided neuronavigation can be improved.

Augmented-reality (AR) is where a device, usually a mobile phone, can overlay digital content onto real world content (the world) in order to create a scene for the user which is more interactive than real life. Mixed-reality (MR) is where these overlaid digital images are viewed through the use of glasses (van Doormaal et al., 2019) worn by the user creating a more immersive experience when compared to AR, as the user is less aware of the device. As the ubiquitousness of AR and MR technology increases in the world in general (Rauschnabel 2020), it is interesting to understand how this relatively new technology can assist society in other contexts such as healthcare which has been recognised for years (Shuhaiber 2004). AR devices which enable MR, such as the Microsoft Hololens, have become a real option for future use in a clinical environment. Not only are they relatively inexpensive in comparison (£3,349-£4,349) with existing technology, they enable the surgeon to view patient information through the Hololens, which overlays this holographic information around or over the surgical field. The possibilities for patients' preoperative MRI or CT scan information to be transformed into accurate 3D renders of patient anatomy, for example the brain, and then to be overlaid via the Hololens exactly over the patient in real life is of great interest to the healthcare and HCI community. This mixed reality technology could now be used to preoperatively plan patient surgery as well as to intraoperatively navigate whilst avoiding important human structures, hopefully reducing the risk of negative patient outcome and working towards minimally invasive surgery (MIS) (Basil & Wang, 2019). Another benefit of a Hololens-based AR/MR system is the ability to use hand gestures and voice commands to navigate the Hololens user interface in a hands free and sterile manner.

However, these new workflows and user experiences have to be researched and tested before they can be given to surgical teams around the world. As mentioned before, accuracy of any new system has to be paramount so that surgical accuracy, and therefore outcome, is improved upon compared to existing methods. Research into accuracy of the system, user interface and user experience is again important for researchers in the HCI community. Systems can be iteratively improved to the point of excellent usability for surgeons but most importantly, can be improved iteratively to achieve clinical accuracy and a new gold standard of navigation in surgery. Having said all that, this leads us to my research question – what are the recent uses of Microsoft Hololens and augmented-reality in neurosurgery since 2018?

2 METHODS

2.1 Selection Method

As I am only interested specifically in the Microsoft Hololens, from now on only to be referred to as Hololens, in the field of neurosurgery, my search terms and review are very focused. The Hololens was chosen because of its technical specifications but also its relatively low cost. Only papers written in English and dated after 2018 were included. A paper chase was conducted to collect a small amount of extra papers that escaped the keyword search. Papers inaccessible to the University of York were also excluded from the literature review. Papers deemed of poor quality were again excluded from the review.

2.2 Keywords and Search Terms

The keywords I used for the literature review were as follows:
"Hololens" AND "Augmented Reality" OR "Mixed Reality" AND "Neurosurgery". These were applied on Scopus, Google Scholar and PubMed.

3 RESULTS AND DISCUSSION

3.1 Accuracy of the System

This was the main motivation (N = 16, 80%) for the studies in the review. As discussed in the introduction, it was likely that this was going to be the case. Within the scope of accuracy however, there were two main focuses. The first being the accuracy of registration of the Hololens-based AR/MR system and the second, the accuracy of navigation of the surgeon using the system.

3.1.1 *Accuracy of Registration of the System.*

This category was quite prominent (N = 3, 15%), because initial registration of the hologram to patient is very important – if the hologram of patient anatomy that is viewed through the Hololens is too far out of register from the patient's anatomy, the surgeon isn't commencing surgery from an accurate starting point which could have disastrous consequences.

The first study (Fick et al., 2021) assesses the registration of a holographic image to the patient and tracking it accurately after the surgical bed is moved. Fiducial Registration Error (FRE) was not clinically viable, although the system tracked the hologram after bed movement. Construct validity is ok, the manipulations and measurements of accuracy of initial registration (probe and anatomical landmark) are used due to current lack of gold standard of hologram registration. Post bed movement accuracy was measured by eye which is not great. Internal validity should be better, low lighting, as acknowledged, could affect the results. The third participant test (33% of sample) involved a different anatomical landmark placement, which is not ideal and resulted in more accurate initial registration. We can generalise to this system in these conditions, until a more robust study is done. Ecological validity wise, the registration experiment was conducted on real patients so this is excellent, although the system wasn't used as the registration method for the actual procedure afterwards, meaning there was no pressure of real surgery on participants. The limitations are sample size, post bed movement tracking was observational and open to bias, initial registration didn't use the most accurate markers which is consistent with current research. Adhesive fiducial markers will be used in future research. All this is noted. Overall, validity, rigour and study quality is ok, more could have been done.

The second study (Kunz et al., 2020), compares the registration and tracking of Infrared Marker (IM) accuracy of two types of tracking systems; 1) sensor dependent short throw reflectivity (str) and short throw depth (std) and 2) being the left front (lf) and right front (rf) environmental cameras. Both setups are clinically viable with str and std M = 0.76mm, and lf and rf M = 0.69mm. Although lf and rf are more accurate, this system requires more equipment so str and std is most feasible. Construct validity is good, measures and manipulations are suitable (IF markers are a well established method of tracking) to assess accuracy with bed movement tracked. Internal validity is good with the same equipment used and same distance to the phantom. A control group with a traditional registration method could have been used however. External validity is poor, we can generalise results to this system, likely even in surgery although it has not been tested. Ecological validity is low because of the phantom, lack of surgical setting and it could be that a breathing patient is harder to register and track. More testing is needed for more robust findings. The researchers claim that this is a system to track IM in a surgical setting, error is within clinically acceptable ranges but more needs to be done to achieve statistically significant findings. No limitations or further research was suggested but comparing both methods against a traditional method of registration would strengthen findings. The study is a good one, both valid and rigorous.

The third study (van Doormaal et al., 2019) compares the accuracy of registration of a Hololens-based neuronavigation system (HN) with a conventional neuronavigation system (CN), on a phantom head and on three patients. Results show that on the phantom, Fiducial Registration Error (FRE) with HN was M = 7.2mm compared with CN M = 1.9mm. On the three patients, HN FRE was M

= 4.4mm compared with CN M = 3.6mm. Hologram drift range was 1-2mm. HN was not clinically accurate compared to CN although HN accuracy increased on patients. Construct validity is excellent with fiducial markers and manual fiducial registration used from previous literature. The comparison with the existing registration system is fantastic. Internal validity is good, it's possible that the MRI method of accuracy measurement for patients compared to CT for phantoms could affect results. Surgeons' familiarity with CN and lack of familiarity with HN could also bias findings but this is acceptable. External validity is poor as researchers can only generalise to this system. Ecological validity is good, a small number of patients are used although it does not use HN to actually guide the operation. No study limitations are mentioned, however, small participant numbers and test runs are obvious limitations. Further study into increased registration accuracy, drift elimination and the effect of objects in the room moving affecting registration is stated. Researchers compare results with previous research which is great. This is a good study.

3.1.2 *Accuracy of Navigation.*

This category (N = 13, 65%) highlights the concerns mentioned in the introduction about the paramount need for system accuracy. The papers below show that without a doubt, the main concern of any Hololens-based neuronavigation system is the ongoing accuracy if it is to be considered an accurate and trustworthy tool in the surgeon's toolkit.

The first study (Gibby et al., 2020) compares an Hololens-based navigation system on patients with a control group on a phantom skull. Results show that comparing target error between groups shows no significant difference (1.73mm less accurate in the patient group) meaning target acquisition was not really affected by patient movement intraoperatively. Target acquisition was not affected by target size or distance to target. Reregistering was not needed 82% of the time. The goal of the study is clear, the manipulations and measurements assess accuracy of navigation so construct validity is excellent. Internal validity is good as a control group is used and the researchers try to control bias by swapping duties of the surgeons. External validity is good. The AR system is tested on real patients although not a large sample. Interestingly, the AR system enabled successful outcomes compared to prior failed attempts via traditional methods. In six other challenging cases, the AR system aided successful navigation to the target site, therefore we can generalise to the potential of this system in a real surgical situation. Ecological validity is high as the navigation system has been tried successfully on real patients in a real surgical environment. The study doesn't mention any further study or limitations, although participant numbers are limited. Other limitations are anatomical differences between procedures of both groups and so cannot be directly compared. However, I think the study is excellent and of high validity and rigour.

Secondly, (Incekara et al., 2018) compared localising a brain tumour borders with the Hololens and with the gold standard of neuronavigation. The AR system was identically accurate to traditional neuronavigation (36% of cases) and became more accurate in the second half of the study indicating accuracy increases with experience. The system isn't clinically accurate for intraoperative surgery, but could be used for preoperative planning. Construct validity was good, the aims of the study are clear and the Hololens system was compared to the gold standard technique. The demographics of patients is varied but skewed towards patients with superficial tumours so internal validity could be better. The study can be generalised to this procedure in the wild due to the ecologically valid nature of the study being a real procedure performed on real patients. Limitations include using a non validated method of outlining the tumour, tumours were mostly superficial, although the deeper tumours were accurately localised too. There could have been a selection bias based on cases that were easier to localise the tumour being more suitable to include in the study. Data collection and analysis is appropriate so this is a good study that is both valid and rigorous.

Next (Ivan et al., 2021) compared the accuracy of tumour outlines via a Hololens-based navigation system versus a monitor and wand-based neuronavigation system (MWBNS). Qualitatively, the Hololens is comparable to MWBNS with 80% excellence accuracy in the second half of the study and 16.7% in the first half indicating a learning curve, confirmed quantitatively with statistical significance between percentage overlap and patient number (p = 0.015). The Hololens is not accurate enough for intraoperative surgery but could be used for planning. Construct validity is good with qualitative and quantitative measures and a real procedure conducted. No training was given so findings could be affected by differing levels of tech experience between surgeons. However, surgeons were blinded to the colours of tumour outlines therefore controlling for bias and increasing internal validity. External validity is good, we can generalise to this procedure and system as the localisation of the tumour is a real procedure. Ecologically this is a good study, the AR system was tested on real patients, in a surgical environment although the surgeons knew they weren't going to use the Hololens for the actual procedure. Although the study is limited in its sample size which is noted, this study has good validity and is rigorous so it is an excellent study. It is stated that consent was not needed as patient care wasn't affected. Clearly the Hololens added time to surgery and therefore the quality of the study is reduced to good.

The next study (Li et al., 2018) compares Hololens guided navigation of external ventricular drain (EVD) insertion and freehand EVD insertion. The Hololens system reduces the number of passes (1.07 compared to 2.33, p < 0.01) and the mean target deviation was less in the Hololens system (4.34mm compared with 11.26mm, p < 0.01). Construct validity is good with clear study aims and target deviation, number of passes and completion time all good measures of accuracy. Internal validity is good, the same equipment is used and both groups are demographically statistically significantly similar, although only one surgeon was used, possibly biassing findings. Therefore, we can't generalise too much. Positively affecting external and ecological validity however, is that the EVD is a real procedure and was conducted on real patients in a real pressured environment, which is also compared to the freehand control group. Data collection and analysis was done well and with validity high and results are compared to previous research and other commercial navigation systems. Limitations and further study addressing low participant numbers and manual registration of the system are noted. Overall, this study is excellent.

Furthermore (Liebmann et al., 2019) conduct a lab experiment to determine the accuracy of the Hololens navigation system on phantom spinal columns. Screw insertion accuracy was M = 2.77mm and trajectory error was M = 3.38° and accurate for clinical use. Construct validity is excellent. Measures and manipulations, which are also based on the surgical procedure, are suitable for measuring accuracy and the study goals are clear. Internal validity is ok, however the phantoms are fixed to the table, unlike in real surgery so this could make it easier to localise and drill. There is also some hologram drift during the experiment. External validity is low because of the phantom spine so the experiment just shows accuracy in a controlled experiment. Ecological validity is similarly low as this is a controlled experiment on a phantom with no surrounding tissue to navigate through and no pressures of real surgery on real patients. There's a potential bias of only having one surgeon perform the experiment and having only 10 attempts at the test. Overall validity is ok mainly because of construct, the results have been compared to existing research which is great. Limitations of lack of tissue around the phantom, the phantoms being fixed to the table and the hologram drift were noted. Future studies on human cadavers and comparison to freehand navigation were also noted. Overall this is a good solid study.

The next study (Müller et al., 2020) compares the accuracy of a Hololens-based navigation system to state-of-the-art pose tracking system (PTS) in a pedicle screw procedure. 3D translational error (TE) of the AR system was M = 3.4mm compared with 3.2mm (p = 0.85) and the 3D angular error (AE) was M = 4.3° compared with 3.5° (p = 0.30) so no statistically significant difference between systems. Procedure time was M = 57.5 seconds in the Hololens group and M = 45 seconds in the PTS group (p = 0.30). Construct validity is good with measures of accuracy taken from Liebmann et al., (2019) and compared with gold standard PTS. Internal validity is ok, there should have been equal comparison groups, probably due to the availability of the cadavers. External validity is again ok, we can generalise to accuracy on a human cadaver with this system. Ecological validity is quite high because of the human spine, however, this was still not equivalent to a real surgical procedure on real patients. One surgeon took part with only 30 attempts in total so there could be a bias here. The surgeon found the AR system intuitive and similar to PTS in other aspects in a post procedure survey. The researchers compare results with previous literature on 2D metrics of accuracy and navigation times which is fantastic. Limitations noted were lack of a real surgical environment or real patient, software issues, jittering of the hologram and ergonomic issues all have to be further researched. The study is valid and rigorous and is an excellent study.

Next (Gibby et al., 2019) assess the accuracy of a Hololens-based system during pedicle screw placement on a phantom spine. Initial registration accuracy was M = 2.5mm radius, needle accuracy = 97% and would have remained within the pedicles. Needles placed approximated M = 4.69mm in the mediolateral direction and M = 4.48mm in the craniocaudal direction from the pedicle bone edge. Time to completion was M = 200 seconds. Construct validity is good, the aims are clear and measures and manipulations test accuracy of the crucial parts of pedicle screw placement, the entry point and trajectory of entry. Internal validity is good, the same registration technique and equipment is used, the training given is the same and the varied skill levels of surgeons enable a better quality of data, although demographics aren't known. We can generalise the accuracy of this system on a phantom. Ecological validity is low due to the phantom and there are no pressures or conditions of surgery. Experienced surgeons could be either overly resistant or receptive to the AR system causing a bias. Limitations are the manual registration of the hologram to the phantom and the silicone not being representative of human tissue, which are noted, and small sample numbers and insertion marks and unrealistic nature of the study, which isn't noted. Future studies into automated registration are noted and the potential of the system to be used in training is a sound one. The study is generally valid, rigorous and overall is a good one.

The next study (Meulstee et al., 2019) assesses the cost of adding the Hololens to an existing neuronavigation system to solve the task switching issue of glancing at an adjacent screen during surgery. Adding the Hololens decreases accuracy overall, M = 1.6mm. Construct validity is good as the aims are clear and the tight-fit and loose-fit tests used to isolate the effects of the Hololens on the system are appropriate. Internal validity is good as interobserver variability is not statistically significant. External validity is ok, you can generalise to this set up. Ecological validity is low as the Hololens is not tested in real surgical conditions on a patient, it's just an experiment with apparatus. Data gathering and analysis is appropriate although more participants are needed. The researchers generalise to a wide range of applications which is a little hasty, more ecologically valid experiments are needed. The findings are noted as being similar to previous literature. Limitations are stated by the authors in the form of further research such as manual registration of the system to improve AR visualisation error, having to add a 4Hz filter to reduce jitter of the hologram and having the hologram more opaque as to not obscure real life. Overall this is a good study.

Next (Schneider et al., 2021) study how a Hololens-based system performs in terms of accuracy during the placement of an external ventricular drain (EVD) on a phantom skull. Ventriculostomy success rate was 68.2%, registration accuracy was M = 2.71mm, tracking accuracy error was M = 0.31mm, system accuracy was M = 3.06mm, overall hit rate was 68.2%, second attempt hit rate was 93.3% showing a learning curve is needed, error when hit was M = 5.2mm with second attempt M = 3.9mm. Participants found the system convenient to use (median Likert Scale score = 4). Construct validity is mostly excellent with tracking marker technique used from previous research, a real procedure task performed and appropriate measures to determine accuracy. The three item survey was basic and not a valid way to measure system satisfaction. Internal validity is ok, the same equipment and training video was used and the surgeons were blind to five ventricular systems beforehand. However, the increased resistance of a hit could affect the final hit position and visible previous insertion holes could potentially affect results. We can only generalise the accuracy of this system. It's not a very ecologically valid study even though a real procedure is tested, it's performed on a phantom and not real surgery. Challenging cases performed on deformed ventricles adds a little validity though. An array of experienced surgeons were used which reduces bias and helps rigour. Limitations of previous insertion holes being visible, only 11 participants used, and a lack of realism in phantom skull haptic feedback are all noted. Future study directions are noted such as tests on cadavers and further researching error reduction. Results were in line with previous research so overall this is a rigorous and valid study.

(Kalavakonda et al., 2019) assess the accuracy and effectiveness of an image rendering algorithm with the Hololens. Results show the algorithm reduces vertex points in the 3D mesh of an object, a skull for example, to speed up rendering through the Hololens but still includes enough critical information to perform a procedure. Construct validity is ok but could be better as the vertex count correlates with time the 3D object is rendered. However, it would be more robust and more accurate to actually time rendering speeds at various levels of detail. Internal validity is therefore affected as the effectiveness of the algorithm in terms of detail is subjective and open to bias as the balance between speed of render and detail is adjusted by the participant. External validity is ok, we can generalise to this system and a little further as the 3D shape could be any body part but with seemingly just one participant further research must be done. Ecological validity is low as the experiment was conducted on a phantom and not a real procedure on a real patient. No limitations are mentioned, although further study with higher specification of computer to enable more detailed 3D shapes or a decrease in render times, a different algorithm and experiments done on a human cadaver are noted. The study is ok but should be better.

The next study (Sun et al., 2020), assesses the accuracy of a Hololens-based catheter tracking system to track inside the skull. Stability of algorithm error was M = 0.33mm, accuracy on a 2D grid error was M = 0.58mm, accuracy on a 3D grid overall was M = 0.85mm and latency of the system was 72ms + 22ms of HTC VIVE. The researchers successfully present an accurate system. Construct validity is excellent as all above manipulations and measures were informed by previous research. Internal validity was generally good, results potentially being affected by wifi signal dips. We can generalise to this system only due to the controlled nature of the study. Ecological validity is low as the tests weren't carried out in surgical conditions on real patients. Apart from the test for algorithm stability it appears the tests were only carried out once. The system is accurate and backed by previous research but I'd like to see more study done on human cadavers to test the robustness of the findings. The researchers don't compare findings within previous literature but overall this is an excellent study mainly because of construct validity and rigour.

This study (Urakov et al., 2019) compares a Hololens-based system to a fluoroscopy-assisted navigation system on a human cadaver with a pedicle screw insertion procedure. The AR system was overall less accurate than the fluoroscopy-assisted navigation system with 7 of the 19 screws completely out of the pedicle which would be devastating for a real patient. Construct validity is excellent as it's a real procedure measured appropriately with gradings (0-3) of accuracy. Internal validity is ok but it's possible that AR initial registration issues meaning the surgeon didn't start from the correct starting point, lack of familiarity with AR and using only one surgeon skewed the AR results. External validity is good. Because a human cadaver is used the results can be generalised to this setup which whilst not the same as actual surgery, really motivates further study on patients. Ecological validity is good because of the cadaver and real procedure used, but it's not a real patient with real pressures of surgery. Results aren't compared to existing literature which is disappointing. Limitations of a single surgeon and poor anterior bone quality are noted however which is great. This is a good study let down by internal validity.

The last study (Van Gestel et al., 2021) compares a Hololens-based system to a freehand navigation system for external ventricular drain insertion (EVD) on a phantom in novice medical students. The AR system significantly improved accuracy (M = 11.9mm untrained and 12.2mm trained) compared to freehand (M = 19.9mm untrained and 13.5mm trained). Both AR error results were similar indicating a small learning curve compared to the freehand technique. Construct validity was good, accuracy measurements of euclidean distance from planned to performed endpoints are appropriate and participants with no prior EVD or AR experience were included to assess the system rather than surgeon skill level. The post experiment survey was adapted from previous research which is good, although there is no mention of the survey design. Internal validity was good as all participants were randomly allocated a group and demographically the groups were evenly spread which is excellent. External validity is good as the researchers assessed the system with novices yielding encouraging results that could apply more generally. Ecological validity is ok, a phantom skull is used which isn't real surgery so the experiment is not very realistic and the resistance from EVD placement was not mimicked, which is used in real life to assist the surgeon to their target. This is noted in the limitations. Appropriate statistical analysis is used with results statistically significant. Qualitative results are described in line with previous research, although only after single pass procedure not multiple as in the literature, which is still excellent. The study champions the use of infrared tracking without external cameras, challenging previous research. This study is an excellent one, both valid and rigorous.

3.2 Evaluation of User Experience (UX) or User Interface (UI) of the System

This category (N = 4, 20%) is primarily focused on testing the user experience of a Hololens-based system. Understandably, usability concerns from surgeons is of great importance and anything that can be designed into (or out of) the system to increase ease of use or reduce task-switching costs is worth researching and implementing.

The first study (Cartucho et al., 2020) evaluates user experience of various preoperative displayed information and interactive functionalities. Surgeons prefer 3D anatomical structures (M = 4.1), Drag & Drop (M = 4.1) and scale and orientation of virtual objects (M = 3.9). Surgeons found the system intuitive and reported potential for use in other fields as well as neurosurgery (89%). Construct validity is ok but could be better. Real patient data and imaging is used to create a real feel to the experiment, however I'd like to see satisfaction measured by a valid scale but feedback is ok. Internal validity is low as 44% of participants were from a neurology background so this could skew findings. All answers in the 5-point Likert Scales were designed negative to positive from left to right which could facilitate an acquiescence bias. We can't generalise further than this system and researchers generalise a little too much by stating the system can be used in neurosurgery, more research needs to be done. Ecological validity is low as the evaluation didn't occur during a real task in real surgery. Limitations such as low participant numbers, 44% of surgeons being neurosurgeons and only short-term testing has been done are all noted. The survey is also self report which is always open to bias. Overall, the study is ok but should have been better.

(Kubben & Sinlae 2019) evaluate user experience of the Hololens with tasks under general theatre and operating lighting conditions, using bright and dark surgical gloves and evaluating the voice recognition with typical background noise. Wearing comfort was ok both with and without regular glasses. Hand gestures were accurate in both lighting conditions with both types of surgical gloves, voice recognition was accurate and the holograms were clear if brightness was set to high. Construct validity is ok. On one hand the tasks, although basic and brief, are representative of tasks during surgery and the study aims are also clear. On the other hand, self-report measures encouraging bias lets construct validity down. Internal validity is poor because of potential self-report bias and only having two participants to evaluate the system. External validity wise, we can generalise that the Hololens is usable and has potential to be used in theatre which is all the researchers claim. Ecological validity is ok as the Hololens has been evaluated in theatre conditions although not in actual surgery with the pressures that come with this. The researchers note the limitations of the evaluation just being 30 minutes. This is an ok study, which could have been improved with more participants, a longer evaluation during a real surgical procedure.

An evaluation of a Hololens-based system for neuronavigation was conducted by (Zhang et al., 2019). The AR system was evaluated for comfort, clarity, manoeuvrability, accuracy and if MR devices could enhance neurosurgeons' understanding of tumour location and related structures. In each category, 6 of 8 cases (75%) said that the system was useful and partly useful the remaining 2 of 8 cases (25%). Construct validity is good as the aims are clear but a validated scale could have been used to measure satisfaction. The manipulations of the experiment are good as it's a real surgical procedure with qualitative data gathered post surgery. Internal validity is poor. Self-report on a non-validated scale is not great and introduces bias. It's not clear how many surgeons conducted the procedures so confirmation bias could be involved if only one surgeon was used. External validation is poor, we can generalise to this system but no more. Ecological validity is excellent as it's a real procedure, on real patients in a surgical environment. Data collection could have been better with more participants and more questions asked with a validated scale. Limitations noted are small sample size and it's only one instance in one hospital. Further research is suggested before clinical use and results are linked to previous research. However, they over generalise stating they have evaluated the system but they haven't yet. Overall, the study is not rigorous enough. It's ok but it should have been better.

The next study (Morales Mojica et al., 2021) evaluates a Hololens-based planning system. The initial ease of use of the system varied between participants, with more training suggested. Compared to desktop-based volume rendering, 3D anatomical structures were easier to understand, made detection and collision avoidance of critical structures easier and enabled easy selection of a trajectory during stereotactic planning. Construct validity is good but could be better, the aims are clear, the task list is iterated for relevance. The measures of evaluation involve no validated scale with just three answer choices and the self-report nature could introduce bias reducing validity. Internal validity is ok but the small sample size hinders good data collection. External validity is just ok as the sample size and task list are small, if both were larger then the system could have been tested more robustly. Ecological validity is low as the evaluation isn't in the context of real surgical conditions, although the task list was based on actual clinical procedures. The results aren't published so we can't comment on thoroughness. Limitations are mentioned such as small study sample, lack of quantitative data and lack of ecological validity. Future study is also mentioned and so is generally ok but could be more valid and rigorous.

3.3 Experiment Methods

3.3.1 *Quantitative Methods.*

This was the most prevalent method (N = 12, 60%). When mixed methods studies are also taken into account, this is (N = 16, 80%).

3.3.2 *Qualitative Methods.*

There were very few studies with qualitative methods (N = 4, 20%). Again, when mixed methods studies are also taken into account, this is (N = 8, 40%).

3.3.3 *Mixed Methods.*

There were very few studies that used mixed methods (N = 4, 20%).

3.3.4 *Test Setup, Phantom, Human Cadaver, Patient or Patient and Phantom?*

These studies can be separated into five categories; test setup (N = 6, 30%), phantom (N = 5, 25%), human cadaver (N = 2, 10%), patient (N = 5, 25%) or patient and phantom (N = 2, 10%).

3.3.5 *Controlled Experiments, In the Wild Experiment or Evaluation Study?*

These studies could be broken down into controlled experiments (N = 13, 65%), in the wild experiments (N = 5, 25%) or evaluation studies (N = 2, 10%).

3.4 Study Type and Participant Numbers

The studies fell into two categories, a full study (N = 6, 30%) or a pilot/feasibility study (N = 14, 70%). Participant numbers were all less than 25 with some studies omitting participant numbers.

3.5 Validity Trends

Common validity trends are that construct and internal validity are generally good as 65% of the studies are controlled experiments. Diversely, this has a knock on effect for a reduction in ecological validity across the study sample as you would expect. Because most of the studies are pilot studies (N = 14, 70%), this also has a negative impact on external validity and the ability to generalise results. All studies had small samples (all ≤ 25), coupled with the low ecological validity of the majority of those also being controlled experiments (N = 8), making it hard to generalise.

3.6 Clinical Accuracy Findings

There is a lack of clinical AR accuracy with only six studies deemed accurate enough (N = 6, 30%).

4 CONCLUSIONS

Most studies focused on accuracy (N = 16, 80%), with 13 focused on accuracy of navigation and 3 on accuracy of registration of the system. Pilot studies dominate (N = 14, 70%) indicating AR neuronavigation is still in its infancy. Moreover, only a few experiments involve patients (N = 7, 35%) and only a few are in the wild experiments (N = 5, 25%). It seems quantitative research (N = 12 plus 4 mixed methods) is at the fore, focussing on accuracy in order to prove the validity of the Hololens AR neuronavigation system, with qualitative (N = 4 plus 4 mixed methods) deemed less important seemingly until a gold standard AR neuronavigation system is found which can then be assessed for improved user experience. General quantitative findings indicate that Hololens-based neuronavigation is viable yet still not clinically accurate enough to be reliable with only a few studies indicating clinical system accuracy (N = 6, 30%). Qualitative findings show the user experience of the system is received well by surgeons, although non-valid surveys and self-report biases show data collection can be improved in future work. Further research should focus on finding one main workflow for AR registration and navigation to get to clinical standards, however, efforts should be made to design experiments that thoroughly test Hololens-based systems on real patients during surgery, simultaneously conducting more ecologically valid studies with larger participant numbers (more than 25 participants) to generalise results. As clinical standards of accuracy are met, further qualitative research and user experience iteration can be done on the new AR systems to fine tune experience for surgeons. Therefore, the focus is on answering three questions. One, can a single accurate and reliable registration method be found? Two, can a single accurate navigation method be found? And three, can the user experience of the Hololens-based AR system be fine tuned? All this will attempt to answer the question – can a Hololens-based AR neuronavigation system give us a new gold standard in neuronavigation?

REFERENCES

Basil, G. W., & Wang, M. Y. (2019). Trends in outpatient minimally invasive spine surgery. *Journal of spine surgery (Hong Kong), 5*(Suppl 1), S108–S114. https://doi.org/10.21037/jss.2019.04.17

Carswell, C. M., Clarke, D., & Seales, W. B. (2005). Assessing mental workload during laparoscopic surgery. *Surgical innovation, 12*(1), 80–90. https://doi.org/10.1177/155335060501200112

Cartucho, J., Shapira, D., Ashrafian, H., & Giannarou, S. (2020). Multimodal mixed reality visualisation for intraoperative surgical guidance. *International journal of computer assisted radiology and surgery, 15*(5), 819–826. https://doi.org/10.1007/s11548-020-02165-4

van Doormaal, T., van Doormaal, J., & Mensink, T. (2019). Clinical Accuracy of Holographic Navigation Using Point-Based Registration on Augmented-Reality Glasses. *Operative Neurosurgery (Hagerstown, Md.), 17*(6), 588–593. https://doi.org/10.1093/ons/opz094

Fick, T., van Doormaal, J., Hoving, E. W., Regli, L., & van Doormaal, T. (2021). Holographic patient tracking after bed movement for augmented reality neuronavigation using a head-mounted display. *Acta Neurochirurgica, 163*(4), 879–884. https://doi.org/10.1007/s00701-021-04707-4

Ghazanfar, M. A., Cook, M., Tang, B., Tait, I., & Alijani, A. (2015). The effect of divided attention on novices and experts in laparoscopic task performance. *Surgical Endoscopy, 29*(3), 614–619. https://doi.org/10.1007/s00464-014-3708-2

Gibby, J., Cvetko, S., Javan, R., Parr, R., & Gibby, W. (2020). Use of augmented reality for image-guided spine procedures. *European Spine Journal: Official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 29*(8), 1823–1832. https://doi.org/10.1007/s00586-020-06495-4

Gibby, J. T., Swenson, S. A., Cvetko, S., Rao, R., & Javan, R. (2019). Head-mounted display augmented reality to guide pedicle screw placement utilizing computed tomography. *International Journal of Computer Assisted Radiology and Surgery, 14*(3), 525–535. https://doi.org/10.1007/s11548-018-1814-7

Göras, C., Olin, K., Unbeck, M., Pukk-Härenstam, K., Ehrenberg, A., Tessma, M. K., Nilsson, U., & Ekstedt, M. (2019). Tasks, multitasking and interruptions among the surgical team in an operating room: A prospective observational study. *BMJ Open, 9*(5), e026410. https://doi.org/10.1136/bmjopen-2018-026410

Grimson, W. E., Kikinis, R., Jolesz, F. A., & Black, P. M. (1999). Image-guided surgery. *Scientific American, 280*(6), 62–69. https://doi.org/10.1038/scientificamerican0699-62

Incekara, F., Smits, M., Dirven, C., & Vincent, A. (2018). Clinical Feasibility of a Wearable Mixed-Reality Device in Neurosurgery. *World Neurosurgery, 118*, e422–e427. https://doi.org/10.1016/j.wneu.2018.06.208

Ivan, M. E., Eichberg, D. G., Di, L., Shah, A. H., Luther, E. M., Lu, V. M., Komotar, R. J., & Urakov, T. M. (2021). Augmented reality head-mounted display-based incision planning in cranial neurosurgery: A prospective pilot study. *Neurosurgical Focus, 51*(2), E3. https://doi.org/10.3171/2021.5.FOCUS20735

Kalavakonda, N., Sekhar, L., & Hannaford, B. (2019). *Augmented reality application for aiding tumor resection in skull-base surgery.* 2019 International Symposium on Medical Robotics (ISMR), 2019, 1-6. https://doi.org/10.1109/ISMR.2019.8710203.

Kubben, P.L., & Sinlae, R.S. (2019). Feasibility of using a low-cost head-mounted augmented reality device in the operating room. *Surgical Neurology International. 10*(26). https://doi.org/10.4103/sni.sni_228_18.

Kunz, C., Maurer, P., Kees, F., Henrich, P., Marzi, C., Hlavac, M., Schneider, M., Mathis-Ullrich, F. (2020). Infrared marker tracking with the HoloLens for neurosurgical interventions. *Current Directions in Biomedical Engineering, 6*. https://doi.org/20200027. 10.1515/cdbme-2020-0027.

Li, Y., Chen, X., Wang, N., Zhang, W., Li, D., Zhang, L., Qu, X., Cheng, W., Xu, Y., Chen, W., & Yang, Q. (2018). A wearable mixed-reality holographic computer for guiding external ventricular drain insertion at the bedside. *Journal of Neurosurgery,* 1–8. Advance online publication. https://doi.org/10.3171/2018.4.JNS18124

Liebmann, F., Roner, S., von Atzigen, M., Scaramuzza, D., Sutter, R., Snedeker, J., Farshad, M., & Fürnstahl, P. (2019). Pedicle screw navigation using surface digitization on the Microsoft HoloLens. *International Journal of Computer Assisted Radiology and Surgery, 14*(7), 1157–1165. https://doi.org/10.1007/s11548-019-01973-7

Meulstee, J. W., Nijsink, J., Schreurs, R., Verhamme, L. M., Xi, T., Delye, H., Borstlap, W. A., & Maal, T. (2019). Toward Holographic-Guided Surgery. *Surgical Innovation, 26*(1), 86–94. https://doi.org/10.1177/1553350618799552

Morales Mojica, C. M., Velazco-Garcia, J. D., Pappas, E. P., Birbilis, T. A., Becker, A., Leiss, E. L., Webb, A., Seimenis, I., & Tsekos, N. V. (2021). A holographic augmented reality interface for visualizing of MRI data and planning of neurosurgical procedures. *Journal of Digital Imaging, 34*(4), 1014–1025. https://doi.org/10.1007/s10278-020-00412-3

Müller, F., Roner, S., Liebmann, F., Spirig, J. M., Fürnstahl, P., & Farshad, M. (2020). Augmented reality navigation for spinal pedicle screw instrumentation using intraoperative 3D imaging. *The Spine Journal: Official Journal of the North American Spine Society, 20*(4), 621–628. https://doi.org/10.1016/j.spinee.2019.10.012

Rauschnabel, Philipp. (2020). Augmented Reality is Eating the Real-world! The substitution of physical products by Holograms. *International Journal of Information Management*. 10.1016/j.ijinfomgt.2020.102279.

Rodwin, B. A., Bilan, V. P., Merchant, N. B., Steffens, C. G., Grimshaw, A. A., Bastian, L. A., & Gunderson, C. G. (2020). Rate of Preventable Mortality in Hospitalized Patients: A systematic Review and Meta-analysis. *Journal of General Internal Medicine, 35*(7), 2099–2106. https://doi.org/10.1007/s11606-019-05592-5

Schneider, M., Kunz, C., Pal'a, A., Wirtz, C. R., Mathis-Ullrich, F., & Hlaváč, M. (2021). Augmented reality-assisted ventriculostomy. *Neurosurgical Focus, 50*(1), E16. https://doi.org/10.3171/2020.10.FOCUS20779

Shaikhouni, A., & Elder, J. B. (2012). Computers and neurosurgery. *World Neurosurgery, 78*(5), 392-398. https://doi.org/10.1016/j.wneu.2012.08.020.

Shuhaiber, J., H. (2004). Augmented Reality in Surgery. *Archives of Surgery, 139*(2), 170–174. https://doi.org/10.1001/archsurg.139.2.170

Sun, X., Murthi, S. B., Schwartzbauer, G., & Varshney, A. (2020). High-Precision 5DoF tracking and visualization of catheter placement in EVD of the brain using AR. *ACM Transactions on Computing for Healthcare, 1*. 1-18. https://doi.org/10.1145/3365678.

Teo, C. (2010). The concept of minimally invasive neurosurgery. *Neurosurgery Clinics of North America, 21*(4), 583–v. https://doi.org/10.1016/j.nec.2010.07.001

Thakkar, S. C., Thakkar, R. S., Sirisreetreerux, N., Carrino, J. A., Shafiq, B., & Hasenboehler, E. A. (2017). 2D versus 3D fluoroscopy-based navigation in posterior pelvic fixation: Review of the literature on current technology. *International Journal of Computer Assisted Radiology and Surgery, 12*(1), 69–76. https://doi.org/10.1007/s11548-016-1465-5

Van Den Bos, J., Rustagi, K., Gray, T., Halford, M., Ziemkiewicz, E., & Shreve, J. (2011). The $17.1 billion problem: the annual cost of measurable medical errors. *Health Affairs (Project Hope), 30*(4), 596–603. https://doi.org/10.1377/hlthaff.2011.0084

Van Gestel, F., Frantz, T., Vannerom, C., Verhellen, A., Gallagher, A. G., Elprama, S. A., Jacobs, A., Buyl, R., Bruneau, M., Jansen, B., Vandemeulebroucke, J., Scheerlinck, T., & Duerinck, J. (2021). The effect of augmented reality on the accuracy and learning curve of external ventricular drain placement. *Neurosurgical Focus, 51*(2), E8. https://doi.org/10.3171/2021.5.FOCUS21215

Urakov, T. M., Wang, M. Y., & Levi, A. D. (2019). Workflow Caveats in Augmented Reality-Assisted Pedicle Instrumentation: Cadaver Lab. *World Neurosurgery, 126*, e1449–e1455. https://doi.org/10.1016/j.wneu.2019.03.118

Zhang, Z. Y., Duan, W. C., Chen, R. K., Zhang, F. J., Yu, B., Zhan, Y. B., Li, K., Zhao, H. B., Sun, T., Ji, Y. C., Bai, Y. H., Wang, Y. M., Zhou, J. Q., & Liu, X. Z. (2019). Preliminary application of mxed reality in neurosurgery: Development and evaluation of a new intraoperative procedure. *Journal of Clinical Neuroscience: Official Journal of the Neurosurgical Society of Australasia, 67*, 234–238. https://doi.org/10.1016/j.jocn.2019.05.038