



Use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings across the NHS

Study Period: May 2024 to April 2025

Report Date: July 9th, 2025



This study was delivered by Great Ormond Street Hospital's Data Research, Innovation and Virtual Environments Unit, with funding from NHS England Frontline Digitisation



The Ambient Voice Technology provider and technology partner for this study was TORTUS AI

Version 1.2 – 16/09/2025

Great Ormond Street Hospital DRIVE Unit

Great Ormond Street Hospital (GOSH) is a leader in research and innovation, with a dedicated unit for digital innovation, the GOSH Data Research, Innovation and Virtual Environments (DRIVE) Unit. Established in 2018 following the introduction of the electronic patient record system at the trust, GOSH DRIVE aims to transform the use of data and technology in healthcare to improve outcomes and experience for children and young people and healthcare staff.

With partnerships across the NHS, industry and academia, GOSH DRIVE supports development and early-stage testing of data and technology solutions to be scaled in practice at GOSH and in the wider NHS. The team have expertise in technology, informatics, project and partnership management, healthcare evaluation and outcomes measurement.

<https://www.goshdrive.com/>

GOSH DRIVE AVT Evaluation Project Team

Name	Title / Project Role
Paul Gough	Ambient AI Programme Manager and Partnership & Projects Manager
Alastair Hill	Observation Team Leader, Project Manager, DRIVE
Nick James	Partnership & Projects Manager, DRIVE
Stephen Mathew	Head of Innovation, DRIVE
Geralyn Oldham	Project Evaluation, Quality & Outcomes Manager, DRIVE
Dr. Robert Robinson	Consultant Paediatric Neurologist, Clinical Lead for AI, GOSH
Dr. Shankar Sridharan	Consultant Paediatric Cardiologist, Chief Clinical Information Officer, GOSH, National Clinical Lead Artificial Intelligence, NHSE
Eleanor Sullivan	Communications and Engagement Manager, DRIVE
Prof. Andrew Taylor	Director of Innovation, GOSH DRIVE
Maya Tomlinson	Data Engineer, DRIVE
Prof. Jo Wray	Project Evaluation, Senior Research Fellow (Health Psychology), DRIVE and Centre for Outcomes and Experience Research in Children's Health, Illness and Disability (ORCHID), GOSH

Acknowledgements

GOSH DRIVE - Observation Team (Baseline and AVT Stages):

- Oliver Thrasher
- Olasumbo Soyege
- Herve Lokoho-Odimba
- Ibrahim Yakubu
- Osasumwen Imade
- Anees Rafiq
- Aravinth Venkataramani
- Jafar Uthanam Kandy
- Darren Brown
- Sayma Bhuta
- Baseer Abdel
- Victoria Stevens
- Kevin Jones
- Duncan Shepherd
- Samantha Banks
- Ali Sayed

Collaborating Teams:



Dr Dom Pimenta Co-Founder & Chief Executive Officer
Jasmine Balloch, Chief of Staff, Tortus
The Tortus Team

Tortus AI developed and provided the AVT app used in the ambient voice Phase 4 trial



**York Health
Economics
Consortium**

Ciara Buckley, Research Consultant
Nick Hex, Associate Director

The YHEC provided an independent economic analysis of Emergency Department data

Health Foundation

Josh Keith, Assistant Director of Data Analytics
Malte Gerhold, Director of Innovation and Improvement
Charles Tallack, Director of Research and Analysis
Alison Moulds, Improvement Fellow

The Health Foundation provided early comment on the study design and on-going input throughout the study period.

Participating Sites:

The evaluation of ambient voice technology across diverse clinical areas would not have been possible without the support and engagement of many clinical and non-clinical staff across all project sites. The DRIVE Project team are very grateful for the participation and co-operation of the following NHS organisations.



Kingston and Richmond
NHS Foundation Trust

Kingston Hospital, Galsworthy Road,
Kingston upon Thames, KT2 7QB

Teddington Memorial Hospital,
Hampton Road, Teddington, TW11 0JL



Crosslands Surgery

Crosslands Surgery, 1 Crosslands
Avenue, Norwood Green, Southall,
Middlesex, UB2 5QY



**Great Ormond Street
Hospital for Children**
NHS Foundation Trust

Great Ormond Street Hospital for
Children NHS Foundation Trust, Great
Ormond Street, London WC1N 3JH



North London
NHS Foundation Trust

North London NHS Foundation Trust
(NLFT), St Pancras Hospital, 4 St
Pancras Way London NW1 0PE



University College London Hospitals
NHS Foundation Trust

University College Hospital, 235
Euston Road, London NW1 2BU



St George's University Hospitals
NHS Foundation Trust

St George's Hospital Emergency
Department, Blackshaw Road, Tooting,
London, SW17 0QT



London Ambulance Service
NHS Trust

London Ambulance Service, 220
Waterloo Road, London, SE1 8SD

Contents

Glossary.....	8
1. Executive summary	10
Learnings and broader system impact.....	12
Headline Findings – Ambient Voice Study	13
2. Introduction.....	14
Current context	15
Phase definitions.....	16
Summary of earlier phases	18
Phase 2	18
Phase 3	19
How Phase 3 further informed Phase 4	20
Phase 4.....	20
Hypotheses.....	20
Study design	21
Sites.....	21
Timeline	22
Sample size	23
Governance	24
Study registration and consent processes	24
Study set-up.....	24
3. Quantitative data	30
Introduction	30
Total Time of Session.....	31
Direct Care Percentage	31
Statistical Methods.....	32
Exploratory Analysis.....	33
Statistical Analysis and Results.....	33
Direct Care Percentage	33

Total Time	36
Limitations	37
4. Survey data	38
Methods	38
Patient and parent experience measures (core sites only)	38
Clinician experience measures (core and non-core sites)	38
Results	39
Patient and parent/carer experience measures	39
Clinician experience measures	45
NASA Task Load Index	53
5. Interview Data	61
Methods	61
Results	61
Participants	61
Themes	61
6. Individual case examples	81
Emergency Department (ED)	81
ED - Key Performance Indicator 1	82
ED - Key Performance Indicator 2	82
ED - Key Performance Indicator 3	84
Ambulance Service - Pan-city ambulance service	85
1. Hear and Treat	85
2. Face to Face	86
7. York Health Economic Consortium (YHEC)	89
Ambient Voice Technology in Generating Clinical Capacity: ED Summary Results	89
8. Team Learning and Playbook (TBD)	92
9. NHS T.E.S.T.	101
10. Discussion	104

Summary of results	104
Strengths and Limitations	105
Next steps	106
11. Appendices.....	108
Appendix A – example information leaflet text.....	109
Appendix B – example consent forms.....	113
Appendix C – example surveys.....	115
Appendix D – interview example topic guide.....	116
Appendix E: Additional quantitative (TimeCat) data analysis	117
Exploratory Analysis	117
Model Reasoning and Residual Diagnostics	128
T-test assumption testing	131
Limitations	132
Appendix F – additional survey data analysis	133
Patient Survey Data	133
Parent-Carer Survey Data	134
Clinician Survey Data	135
Appendix G - Sheffield Assessment Instrument for Letters	137
Appendix H – Net Promoter Score	138
Appendix I - The NASA Task Load Index (NASA-TLX)	139
12. References	140

Glossary

Ambient Voice Technology (AVT) is the use of two or more AI models (including a generative AI model) used together to passively, and temporarily, capture a consultation and automatically generate notes and letters from the audio. Specifically, it is the process of converting audio into transcript and then transcript into note - some modern multimodal models can do this in one step now so specifically a large language model is the core requirement.

Akaike information criterion (AIC): Akaike information criterion (AIC) is a number score used for measuring model fit, when compared to other models, in order to determine the best fitting model for a given dataset. A lower score indicates a better fitting model.

Bayesian information criterion (BIC): Bayesian information criterion (BIC) is a number score also used for measuring model fit, when compared to other models, in order to determine the best fitting model for a given dataset. A lower score indicates a better fitting model. BIC penalises model complexity more than AIC, making it useful in scenarios where overfitting could be a concern

Core site: There were five core sites where the following pre- and post-intervention data were collected:

- Recording of in-clinic time-motion variables using the TimeCat app (www.timecat.org)
- Post-consultation clinician survey and semi-structured interview data
- Post-consultation patient or carer survey data

Non-core site: there were 4 non-core sites where the following data pre and post-intervention details were collected:

- Post-consultation clinician survey and semi-structured interview data
- Post-consultation patient or carer survey data

Direct Care: clinician directly conversing with, examining or listening to the patient. This was measured in minutes and seconds at core sites

DRIVE: Digital Research, Innovation and Virtual Environments unit (GOSH DRIVE) aims to transform the use of data and technology in healthcare to improve outcomes and experience for children and young people. DRIVE is a state-of-the-art unit dedicated to innovation through data and digital technologies, with partnerships across the NHS, industry and academia.

EHR/EMR/ePCR: Electronic Health Record / Electronic Medical Record / electronic Patient Care Record

Feature flagging: Feature flagging is a technique which enables software developers to continuously ship updates to a product; the feature enables developers to control who can see which version at any given moment in time. However, this can only happen when the on/off switches (flags) are activated. For the AVT study the flag (switch) was turned off for every site for the duration of the study period meaning that the AVT product remained unchanged throughout the trial period.

Indirect care: clinician activities whilst the patient is present (broken down into the following)

- Computer Notes - note taking via typing or dictaphone (min:sec)
- Computer Orders - orders/ form completion (min:sec)
- Computer Read - reading/analysing charts or scans (min:sec)
- Written Notes - note taking by hand (min:sec)
- Other indirect i.e. telephone appointment or when a colleague is brought in

NASA Scale / Nasa Task Load Index: The tool is a subjective workload assessment tool developed by the Human Performance Group at NASA's Ames Research Center. It is designed to evaluate the perceived workload of individuals performing tasks, particularly in complex and high-demand environments such as aviation, space operations, healthcare, and human-computer interaction studies. The NASA-TLX provides a multidimensional rating of workload based on six subscales (mental, physical and temporal demand, performance, effort and frustration level), allowing researchers or practitioners to assess the mental and physical demands of a task from the participant's perspective.

Net Promoter Score: The Net Promoter Score (NPS) is a metric used to gauge patient / family satisfaction by measuring how likely they are to recommend a healthcare provider / service to others. It is based on a single question: "On a scale of 0 to 10, how likely are you to recommend [healthcare provider / service] to a friend or colleague?"

Platform play: Refers to "A platform approach to Ambient Voice Technology (AVT) does not mean choosing a single product or vendor across the NHS. Instead, it involves adopting a range of assured technologies that meet common standards for clinical safety, cybersecurity, and evidence of benefit. These tools should integrate into a shared data environment, enabling visibility across sites, consistency in delivery, and a joined-up approach to training, governance, and implementation. This avoids fragmentation, reduces duplication, and supports scalable, equitable deployment across the health system. Importantly, this approach creates a unified data estate that allows for system-wide analytics, benchmarking, and deeper secondary insights. By feeding structured data from multiple AVT tools into a common framework, the NHS can track impact, identify variation, and inform future planning. It unlocks better value for money, strengthens operational decision-making, and ensures that the benefits of AVT extend beyond isolated pilots to support lasting, system-wide transformation."

TimeCat: an online time capture tool developed to support data capture for time and motion studies (TMS) was used for TMS data collection ([see Appendix E](#)) for a TimeCat application image). Time-and-motion data were collected by clinic observers who categorised the actions of the clinician into one of seven categories: "Computer Notes", "Computer Orders", "Computer Read", "Written Notes", "Direct Care" (encompassing the time spent speaking to or examining the patient), "No Care/Absent", or "Other Indirect"

SAIL: The [Sheffield Assessment Instrument for Letters](#) is an assessment instrument developed from a consensus framework for good practice in written communication. It comprises an 18-point checklist and a global rating scale. Scoring tool can be seen in [Appendix G](#).

1. Executive summary

Background

The Ambient Voice Technology (AVT) Phase 4 evaluation represents the first scientific, multi-site assessment of AVT across the NHS, designed not simply to test whether the technology works in isolated pockets, but to rigorously determine its true value, usability, and impact across care settings. Unlike fragmented local pilots which often lack power, governance consistency, or strategic follow-through, this programme adopted a unified, NHS-led methodology to generate findings that are robust, scalable, and clinically meaningful.

Purpose and approach

The use of Electronic Health Records (EHRs), whilst bringing the clinical record together in one place, has also created an administrative burden with regard to data entry; patient-facing clinical staff have taken on data entry roles, which may contribute to cognitive burn-out. Furthermore, the experience of users in healthcare has changed from a face-to-face consultation, to one where we talk to the side or back of the clinician (or to a screen), whilst the clinician is focused on typing, reducing the human contact we know is so important for the delivery of high quality, compassionate care. The aim of this evaluation was to assess whether AVT with Generative Artificial Intelligence, which generates clinical notes and letters from natural dialogue, could deliver real benefits for patients, clinicians, and the wider system.

The Ambient Voice Technology was tested across a diverse range of clinical contexts: adult and paediatric care, mental health services, an emergency department, the London Ambulance Service (LAS), a community hospital, and general practice.

To achieve meaningful scale and reliability, the programme deployed a structured methodology that enabled local digital governance with central programme delivery. A dedicated strategic delivery team oversaw the rollout, and core sites followed a controlled, pre-post design with clinicians first observed under standard documentation conditions, and then with AVT in active use. This work led to the inception of the NHS T.E.S.T. framework, a rigorous, vendor-agnostic model, developed during this programme to assess technology safety, efficacy, and

scalability. This approach does not attempt to standardise which technologies are used but provides a comprehensive framework to select the best technology.

Key findings

Quantitative and Clinical Outcomes

- **23.5% increase in direct care:** Median increase from 70.0% (baseline) to 86.5% (AVT) in core sites, freeing clinicians to focus on patients rather than screens.
- **Supporting the workforce:** Cognitive Load -Significant reduction in stress and mental effort, as measured by the NASA Task Load Index.
- **Consultation dynamics:** Enhanced patient-clinician interaction, 8% shorter appointments, and increased focus on clinical care.
- **Clinician experience:** Improved satisfaction with time, attention, and documentation accuracy; the technology was widely viewed as enabling better care.

Operational and Economic Impact

A **13.4% increase in patient capacity per shift** was identified in the Emergency Department.

Results per Individual:

51.7% reduction in documentation time at the individual clinician level. Use of the AVT tool saved an average of 6 minutes per documentation task (12mins pre), equating to 47 minutes saved per shift. This time saving enabled each A&E staff member to see one additional patient per shift.

Results per Trust

Annually, the **potential savings total £1,438,847.19** for reduced documentation time and £5,364,458.78 for additional capacity across the 90 staff within the trust.

Results per England

At the national level, applying the same assumptions* to the full-time equivalent workforce of 11,055 A&E doctors results in significant cost savings. AVT technology could enable.

- a. **9,259 additional A&E attendances per day** with
- b. **National projections of £658 million in clinical capacity gains** and
- c. **£176 million in documentation savings** if scaled across all A&E doctors.

The analysis assumes that 80% of the time saved can be directly reused by clinicians to see additional patients in the A&E department (see full report for assumptions and limitations of modelling National use). These results underscore AVT's ability to boost productivity and improve staff wellbeing, freeing clinicians from repetitive documentation tasks and allowing them to focus on what matters most, patient care.

Learnings and broader system impact

The success of whether AVT enhances how we deliver care across the NHS, does not rest on the technology alone; the study confirmed that effective adoption depends on aligning people, processes, and platforms. Key enablers included accelerating uptake by supporting sites navigate digital governance, clinician enrolment, at-the-elbow training, custom template design, work-flow integration, and clinician feedback loops. With these in place, AVT offered multipolar benefits: improving patient-clinician relationships, enhancing documentation quality, protecting workforce wellbeing, and creating opportunities for service redesign. Beyond immediate outcomes, this work has catalysed system-level transformation. The learnings meaningfully supported the creation of: NHS England's national guidance on AI-enabled ambient scribing (published April 29th). The NHS T.E.S.T. framework has been accepted as a key tool to support the evaluation and selecting AVT solutions across the NHS.

Conclusion

This work has shown that the NHS can scientifically assess and safely scale technology to deliver tangible improvements in care quality, clinician experience, and system efficiency. AVT is not just a tool, it is a capability, and its full value can be leveraged if it is delivered strategically as a 'platform play', not as isolated pilots that focus on the technology alone. Strategic deployment capability is needed. By investing in a coordinated, assured, and evidence-led deployment, the NHS can unlock AVT's full potential to support patients, protect staff, and improve the way we deliver care across the system.

Headline findings are shown in Figure 1.1.

Headline Findings – Ambient Voice Study

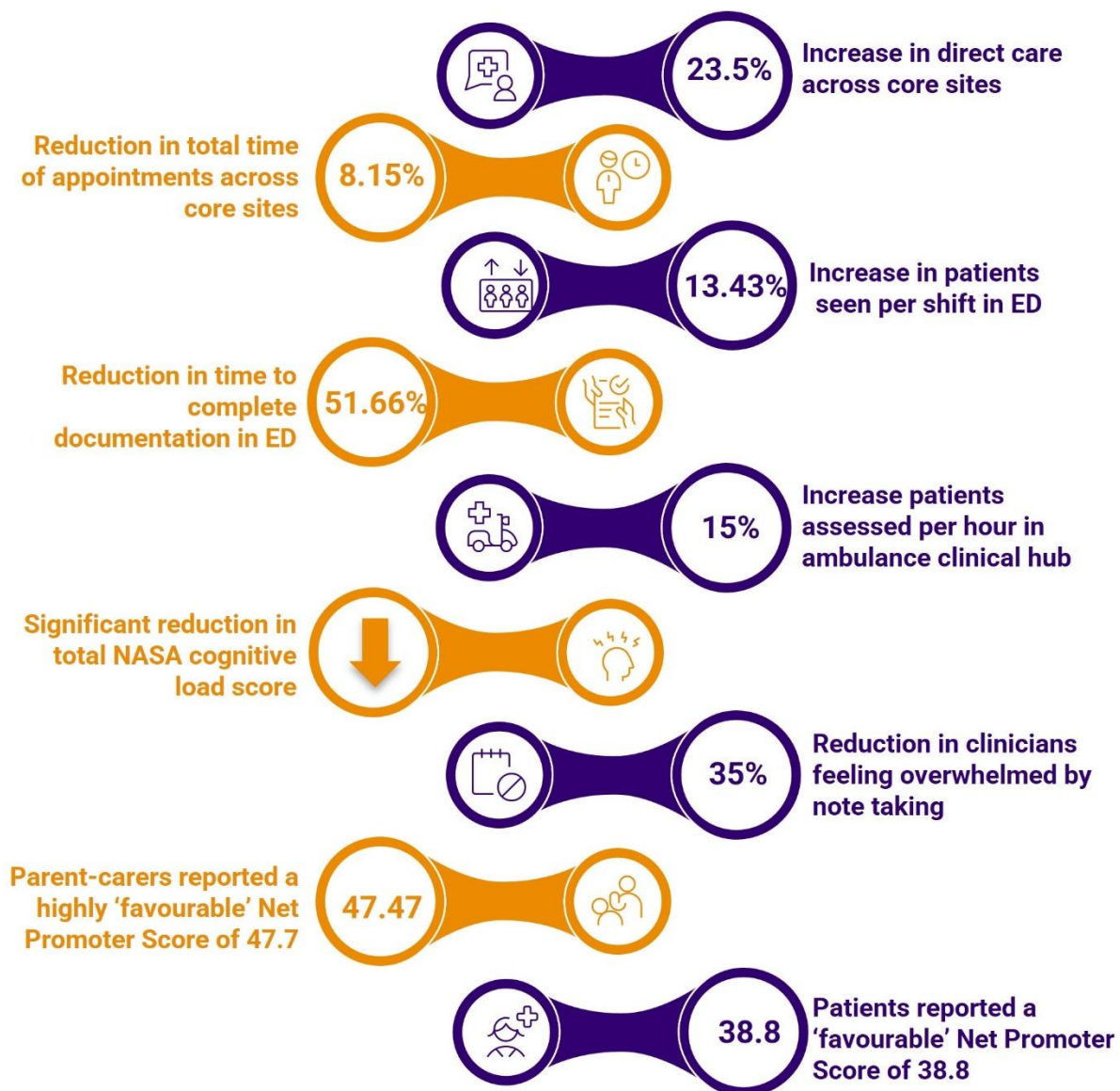


Figure 1.1: Headline Figures - Ambient Voice Technology (AVT) Phase 4 evaluation

2. Introduction

The NHS faces unprecedented challenges, exacerbated by the COVID-19 pandemic, leading to significant waiting lists and delayed care for patients across the UK. In response, the UK Government has proposed increasing NHS appointments by 40,000 per week to alleviate pressure and ensure timely access to healthcare services(1).

This involves a multifaceted approach aimed at reducing waiting lists and enhancing healthcare delivery. Suggested levers to achieve this include extending operating hours and creating shared NHS waiting lists to access appointments. Additionally, NHS staff will receive extra pay for out-of-hours work. A key approach will be the intelligent use of technology, innovation and artificial intelligence to support care and optimise resource use.

NHS services across the board are under significant financial and operational pressure – waiting lists for appointments and treatment are unacceptably long, the number of patients that the NHS is expected to care for is growing, clinicians are over-burdened and the financial pressures under which Trusts are operating are considerable. As a result, the need to deliver timely care more efficiently and effectively is intensifying. Furthermore, clinicians now have to process higher volumes of complex clinical data and meet an ever greater number of regulatory requirements. The consequences are an escalation in the cognitive load clinicians experience and rising rates of mental fatigue and burnout; patient care is impacted and patients' experience of care diminished, putting further pressures on an already overstretched system.

Amongst the demands placed on clinicians, use of electronic health records (EHRs) has been identified as one of the contributory factors to burnout (2, 3). Whilst EHRs create a strong clinical narrative and deep phenotypic record of a patient to support care delivery, the complex structure of these systems and highly structured nature create a large burden of clinical administration that can eat into direct clinical care. Finding remedies is imperative and technology is likely to have at least some of the answers. One such solution is artificial intelligence (AI), the use of which is growing exponentially in health settings, with reported benefits for both clinicians and patients (4). More specifically, the use of ambient AI documentation platforms offers the potential to effectively address some of the administrative burden through efficiently capturing and organising clinical information, thereby streamlining clinical workflows(5), reducing the time that clinicians need to spend interacting with electronic health records, reducing physician workload and maximising direct patient

and family care (6-8). Such platforms are seeing an uptick in adoption in health care settings but the evidence to support their implementation in the UK National Health Service is limited, although AI-generated documents have been suggested to be the '*biggest bet*' to improving NHS productivity (9).

Recent years have seen significant progress in the field of AI and particularly in the deployment of Large Language Models (LLMs), which have demonstrated improved accuracy, speed and utility with each subsequent generation. Such development means that these models are increasingly being deployed to address real world issues in a variety of industries, including healthcare. However, issues continue within the field of LLM development, and in particular hallucinations and omissions persist, albeit decreasingly so. (10).

The NHS Long Term Workforce Plan (10) emphasised the importance for staff retention and the reform of working practices "*to take advantage of new technology that frees up clinicians' time to care*". Unlike other health care systems such as that in the US, the NHS typically does not have medical scribes (health care professionals who assist clinicians by documenting patient information in the electronic health record) and the burden of documentation falls on front-line clinicians. The potential for reduced documentation time to result in improved productivity, reduced staff burnout, improved clinician experience and improved patient outcomes is huge. Furthermore, an increase in the proportion of direct patient care during a consultation is likely to result in improved patient experience and engagement with health care. However, there are still challenges related to governance and diverse clinical workflows in a UK setting and evidence-based proposals for how to roll out the technology at scale are yet to be made.

Current context

In 2023, Great Ormond Street Hospital for Children NHS Foundation Trust (GOSH), an inner city specialist paediatric hospital, entered into an innovation collaboration with a health-technology start-up company to test and evaluate an ambient voice technology (AVT) tool. It was essential for us to standardise the science with a single supplier, ensuring methodological consistency across our deployments. Equally, we needed a partner that would not only align with our strict NHS governance and assurance frameworks but also move at pace—responding rapidly to frontline needs rather than leaving us on a product roadmap. Choosing a UK-based vendor who understood both the technology and the regulatory environment allowed us to set a new benchmark for safe, secure, and scalable AI adoption in the NHS.

We partnered with Tortus AI because they meet the stringent assurance standards we demanded across data protection, cybersecurity, clinical safety, and model governance. They completed multiple independent certifications at our request—including Cyber Essential Plus, bias assessments, and medical device accreditation—demonstrating a commitment to safety and scientific rigour.

The AVT tool was designed to transcribe medical consultations and automatically generate notes and/or clinical letters in a style specified by the clinician. The tool uses templates to determine this style, as well as the content and format of the clinical notes. Clinicians built a personalised template in the AVT tool's template builder – either manually or by entering in an actual clinical note (anonymised) - from which the tool would ascertain the format, style and content. This template could then be applied to all future consultations with the tool.

In light of potential consequences arising from hallucinations and omissions and to mitigate against their impact, the technology we tested is designed to work with human (clinical) oversight and require approval of the output as part of a clinician's professional responsibility. The tool was developed to work alongside EHRs with clinicians accessing the AVT tool through either a secure desktop application or web-based version of the AVT tool. Clinicians were able to incorporate clinic notes and letters generated by the AVT tool into the EHR once they had reviewed and edited the documents to ensure their accuracy and clinical validity.

Phase definitions

To comprehensively assess the AVT technology we adopted a phased methodology borrowed from the principles of a pharmaceutical trial, with increasing exposure in each phase to real patients with decreasing risk, as the technology was proven over time after each stage gate. Phase 1 involved no patients at all, Phase 2 involved real physicians but professional patient-actors, phase 3 was an initial feasibility study with a limited number of real patients, before full roll out in phase 4 in multiple real-world settings. Each phase garnered new learnings and protocol adaptations as per a typical pharmaceutical trial, and each phase therefore informed the next, allowing iterative product development alongside protocol development in parallel. (Figure 3.1).

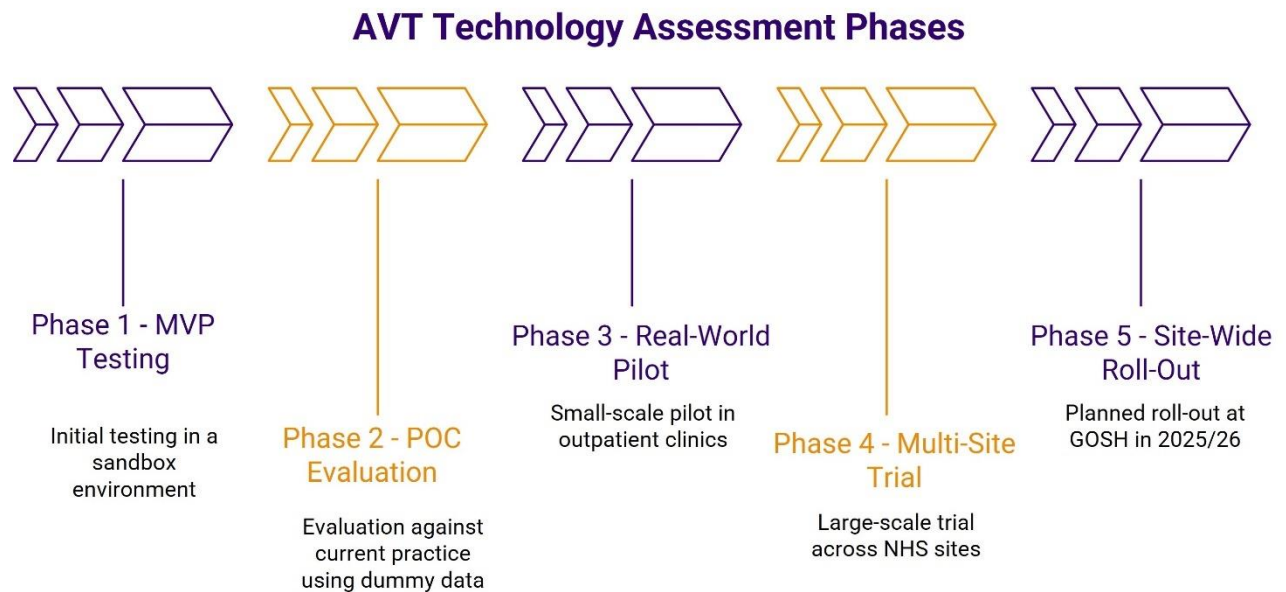


Figure 3.1: Phases of the AAI technology assessment; phase 4 is the subject of the current report

- **Phase 1** — Tested a minimum viable product. This was completed when the AVT tool was tested in a ‘sandbox’ environment at GOSH with the DRIVE team.
- **Phase 2** — evaluated the AVT tool against GOSH’s current practice, using a Proof of Concept (POC) EHR environment. This phase used dummy patient data, professional medical actors and real GOSH clinicians in a simulated clinic environment, using patient and clinician transcripts to support the consultation.
- **Phase 3** — a small-scale real-world pilot evaluation conducted in outpatient clinics at GOSH to assess the AVT tool against current practice, using GOSH’s live EHR environment.
- **Phase 4** – this phase involved a multi-site trial of the AVT tool, designed to evaluate its performance across a broad range of clinical settings within the NHS. The aim was to test the hypothesis that this technology could meaningfully enhance care delivery, demonstrate real clinical utility, and prove scalability across diverse healthcare environments. By assessing the tool in different institutional workflows and gathering extensive feedback from frontline staff and stakeholders, the evaluation sought to generate robust,

generalisable findings. Unlike small-scale or isolated pilots, this panoramic approach was intended to provide a clear, evidence-based assessment of the tool's value to NHS services. The findings from this phase form the basis of this report."

- **Phase 5** – site-wide roll-out at GOSH – planned for 2025/26

Summary of earlier phases

Phase 2

The Phase 2 evaluation of the AVT tool at GOSH was conducted through simulated consultations using a proof of concept EHR environment (11). Eight clinicians performed 48 simulated consultations with professional medical actors and the evaluation compared the current EHR workflow to an adapted workflow with the AVT tool. Consultations were 20 minutes long. The primary objectives were to assess whether the AVT tool could enhance documentation quality, improve clinician and patient experience, and increase operational efficiency.

Results showed improvements in clinical documentation quality, measured using the Sheffield Assessment Instrument for Letters (SAIL) (12). The percentage of clinic notes rated as good or very good rose from 43% to 100% when using the AVT tool, and clinic letter quality improved from 29% to 70%. Clinicians also reported enhanced interactions with patients and families, with 100% agreeing that they could give their full attention during consultations, compared to 66% at baseline (i.e. without the AVT tool). Additionally, operational efficiency was boosted, with a 26.3% reduction in consultation length, therefore saving an average of 3 minutes and 13 seconds per consultation. From a technical perspective, Phase 2 demonstrated that the AVT tool functioned effectively in multi-speaker scenarios and filtered out non-clinical dialogue effectively. At the same time, the Phase 2 evaluation identified areas for improvement ahead of Phase 3. Clinicians expressed the necessity for training on the system in order to improve familiarity with the system during the consultation and to help expedite post-consultation amendments to the AVT output. The study suggested varying consultation durations in future trials and refining the recorded typing time metric to account for context switches between speaking and typing.

Overall, the Phase 2 results marked a significant step forward, providing valuable insights into the potential of AVT tools to enhance clinical efficiency and documentation quality in real-world healthcare settings.

Phase 3

The Phase 3 pilot evaluation of the AVT tool was conducted during live outpatient clinics at GOSH. Written consent was obtained from all participating families (n=98). The objectives were to assess the AVT tool in terms of operational efficiency - evaluating time spent on direct (examining or directly conversing with a patient/carer) versus indirect (reading/writing notes, ordering tests) care; clinician, patient and family experience (through surveys and interviews); and documentation quality (using SAIL to analyse clinic notes and letters for quality). Phase 3 was conducted over 3 months and each of the 11 participating clinicians was observed in two or more clinics, firstly at baseline and secondly using the AVT. Families were provided with information about the pilot study prior to their clinic visit and were asked to either verbally consent (baseline clinics) or provide written consent (AVT clinics) on the day of clinic.

Results indicated a 9% increase in the proportion of time spent providing direct patient care with the AVT compared with baseline (71% of consultation time on direct care vs 80%). There was a 7% reduction in time spent on note-taking with the AVT, although consultation length remained unchanged (20 minutes on average). No significant difference was observed in the turnaround time for generating clinic letters between the AVT and traditional methods.

Clinicians reported improved experiences: there was a 19% increase in the proportion agreeing that they had sufficient time with patients, a 13% increase in the proportion agreeing that they were able to give full attention to patients, and a 15% reduction in the proportion agreeing that computer tasks were distracting.

Interview feedback was mixed, with some clinicians praising the AVT for enhancing patient interactions and reducing administrative tasks, while others expressed frustration with the AVT's limitations, particularly its difficulty in capturing complex medical details and non-Western names. Parents and carers reported a generally positive experience, with 40% noting an improvement in interactions during the AVT phase. Children's and young people's surveys showed modest improvements, including a 12% reduction in perceived computer-related disruptions during consultations.

SAIL analysis showed no consistent improvement in the quality of clinic notes or letters produced with the AVT compared to baseline (which is perhaps not surprising as all letters, baseline and AVT, were checked and authorised by the clinician before they could be sent out). The evaluation of document quality remained resource-intensive and challenging, and there were concerns about potential bias due to the

recognisability of AI-generated documents. Inter-rater reliability of the scoring was very poor.

How Phase 3 further informed Phase 4

Key learning from Phase 3 which informed aspects of Phase 4 included more personalised training and better template configuration. Given the poor inter-rater reliability of the quality scoring of the notes and letters and its time-consuming nature, it was decided not to score the documentation quality due to lack of objectivity but to explore this aspect subjectively through the surveys and interviews.

Phase 4

Aim: To trial and evaluate the performance of ambient voice technology across a broad range of clinical settings within the NHS.

We wanted to assess whether the technology could meaningfully enhance care delivery, demonstrate real clinical utility, and prove scalable across diverse healthcare environments. By assessing the tool in different institutional workflows and gathering extensive feedback from frontline staff and stakeholders, the evaluation sought to generate robust, generalisable findings.

Hypotheses

Overall hypothesis:

The use of ambient voice technology in clinician-led encounters in multiple clinical settings will improve the delivery and quality of care, specifically addressed by the following hypotheses (Figure 3.2):

1. The use of ambient voice technology in an encounter will increase the percentage of direct care delivered during a consultation (vs indirect care).
2. The use of ambient voice technology in an encounter will decrease the total time required to see a patient.
3. The use of ambient voice technology in an encounter will improve clinician experience.
4. The use of ambient voice technology in encounters will improve patient experience.

Exploring the Impact of Ambient Voice Technology



Figure 3.2: Diagrammatic representation of the study hypotheses

Study design

A non-randomised multi-centre within-subject, pre-post intervention trial.

Sites

To ensure adequate power for the evaluation overall, the initial focus was on collecting clinical appointment clinician activity at the core sites, where the methodology was uniform and data collection homogeneous. Five core sites were included, with the aim of recruiting at least 100 clinicians across the sites who, in observed clinics, each saw at least 10 patients in the baseline condition and 10 patients in the AVT condition. The five sites were:

- Crosslands Surgery – general primary care practice

- Kingston Hospital – acute hospital
- University College London Hospital – acute hospital
- Teddington Community Care – community hospital
- Great Ormond Street Hospital – acute hospital

Core sites were able to support deployment of a staff ‘observer’ to assess clinical use of technology at baseline with their existing systems and then with the addition of the AVT tool.

Additional sites were recruited with the aim of capturing a minimum of 5000 AVT encounters overall across a range of diverse patient interactions, with those in the non-core sites not directly observed as their workflows and clinical settings precluded an observer in the clinic room in every appointment. The exception to this was the Primary Care site where observers were present in consultations during the core assessment but not when the tool was assessed to support business as usual practice in everyday use in primary care. Patients in all non-core sites were informed that AVT was being used to support the consultation. The non-core sites were:

- North London Mental Health Partnership (NLMHP) – perinatal outpatients
- St George’s Hospital – emergency department
- London Ambulance Service (LAS) – roadside and clinical hub
- Crosslands Surgery – general primary care practice (Crosslands took part in an additional longitudinal phase, after completing a phase as a Core site)

The remainder of this document uses site ID to confer a degree of anonymity to the findings.

Timeline

Data collection for Phase 4 took place over 12 months, from May 2024 to April 2025. Sites 1 to 3 ran consecutively but an increase in the research team observer pool enabled remaining sites to run concurrently for part of the time (see Table 3.1).

Table 3.1: Timeline for site participation (May 2024 through to April 2025)

Site ID	Core Non-core	May 24	June 24	July 24	Aug 24	Sept 24	Oct 24	Nov 24	Dec 24	Jan 25	Feb 25	Mar 25	Apr 25
Site 1	Core												
Site 2	Core												
Site 3	Core												
Site 4	Core												
Site 5	Core												
Site 6	Non-core												
Site 7	Non-core												
Site 8	Non-core												
Site 9	Non-core												

Sample size

The power calculation to determine sample size was based on one of the primary outcomes – the time spent providing direct care. In order to detect a difference of 10% in direct care, with a power of 90%, 89 clinicians with 10 consultations in each arm (baseline and AVT) would be required. This sample size was also calculated to be sufficient to enable statistical comparison of the primary outcomes of change in total time and change in clinician experience. Allowing for 20% attrition, 112 clinicians needed to be recruited to reach a sample size of 89.

Data collection at core sites (detailed methods are described in chapters 4, 5 and 6).

- Recording of in-clinic time-motion variables
- Post-consultation clinician survey and interview data
- Post-consultation patient and/or carer survey data

In addition, data on the number of clinics, first/follow-up appointment, use of translator and whether the patient was accompanied were collected, together with proportion of patients/parent/carers approached who consented to participate for each clinic.

Data collection at non-core sites

- Post-consultation clinician survey and interview data
- Two non-core sites collected bespoke data about clinician use of the AVT technology related to time and activity

Governance

Use of AI tools in healthcare is advancing at pace. Ensuring that appropriate information governance and ICT controls were in place was essential to successful delivery of the evaluation programme.

The team worked closely with Information Governance and ICT colleagues at each participating site to complete the required documentation and ensure the security of the tool. Tortus provided a webpage with links to all of their governance and security certification. However, carrying out these thorough checks took a significant amount of time and work with clinicians could not commence until this had been completed.

Study registration and consent processes

The study was registered at GOSH as a clinical evaluation study.

Clinicians provided verbal consent for their participation in the trial and consent for surveys was assumed if a completed survey was submitted. Site-specific information leaflets were provided to patients and/or parents/carers. Patients/parents/carers at core sites were asked to verbally consent for observers to be present at baseline and to provide written consent for use of the AVT. Consent was not taken at non-core sites but all participants were informed about the AVT and asked to verbally consent to its use during their consultation.

Study set-up

A [playbook](#) for study set up at each site is provided in chapter 9, including approaches to site engagement and sign up. This was a key aspect to success, enabling local teams to navigate full and complete NHS digital governance to support deployment. A protocol for the study to guide processes at each site was written prior to recruitment of the first site, with clear roles and responsibilities for both the start-up company and members of GOSH outlined.

The product

The AVT product was being continuously improved throughout the study; each site was signed up to the latest build of the product and then locked to that version for the duration of their involvement, which was controlled centrally via feature flagging (see [glossary](#)).

Study team training

The GOSH study team were trained in the use of the AVT, how to revise templates and troubleshoot any technical issues. They were also given TimeCat (13). training if they were going to be part of the team observing clinics.

Clinician training

Each clinician in the core sites received training in the use of the AVT technology and templates prior to using the AVT in clinics. Any clinician in core sites who had not completed training was not included in the study. Some clinicians at non-core sites did not receive formal training. Each clinician set up an account with the AVT company to give them access to the software. Clinicians were encouraged to familiarise themselves with the AVT and templates prior to the study.

Core site set up

Following site approvals, the site lead for the project was responsible for recruiting local clinicians and providing their details to the study team at GOSH together with clinic schedules (days of the week, time of clinic, location of clinic) for those clinicians who signed up. Site-specific information leaflets were developed for patients and parents/carers in which the AVT technology and the trial were explained together with the fact that use of the AVT would not result in any changes to the appointment. It was made clear to potential participants that they were under no obligation to participate and could ask for the observer to leave the room at any time or stop participating without giving a reason and without their medical care being affected.

Baseline condition

Prior to each appointment, the study team observers were responsible for obtaining verbal consent from patients/parents for the observer to be in the clinic room during their appointment. Observers completed the TimeCat recordings during the appointment and at the end of the clinic appointment invited the patient/parent to complete a Patient Reported Experience Measure - PREM (on the GOSH SmartSurvey platform) either using an iPad provided by the observers or via a QR Code (the latter option was rarely used). Paper copies were also available for completion by children and young people. At the end of the clinic observers also invited clinicians to complete a clinician experience measure either on an iPad or via a QR code ([Appendix C](#)).

AVT condition

Once individual clinicians had seen at least 9 patients in the baseline condition they were provided with the AVT software (installed on a clinic computer or laptop) and the necessary microphones so that the AVT technology could be used in the subsequent consultations. The study team observers undertook the same roles as for the baseline condition with the exception of consent from the patients/parents, which was written rather than verbal and included consent for the use of the AVT technology during their consultation. The experience surveys were broadly similar to those for the baseline condition with the addition of some specific questions related to the use of the AVT technology (Appendix C).

Non-core site set up

Non-core site set up was similar to that for core sites but TimeCat data were not recorded and there was no clinic observation or formal consent process. Prior to using the AVT technology in the clinic setting, participating clinicians were emailed a link and QR code to complete a baseline survey. Once they had completed clinics using the AVT they were emailed a second link and QR code to complete the AVT survey. For some sites additional questions were included which were specific to their site – for example, related to the emergency department or ambulance service. No patient or parent/carer PREM surveys were used in non-core sites due to absence of clinic observers to administer them.

Figure 3.3 illustrates the evolution of the study over time in terms of training method, software versions and data collection period for core and non-core sites.

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings
Across the NHS

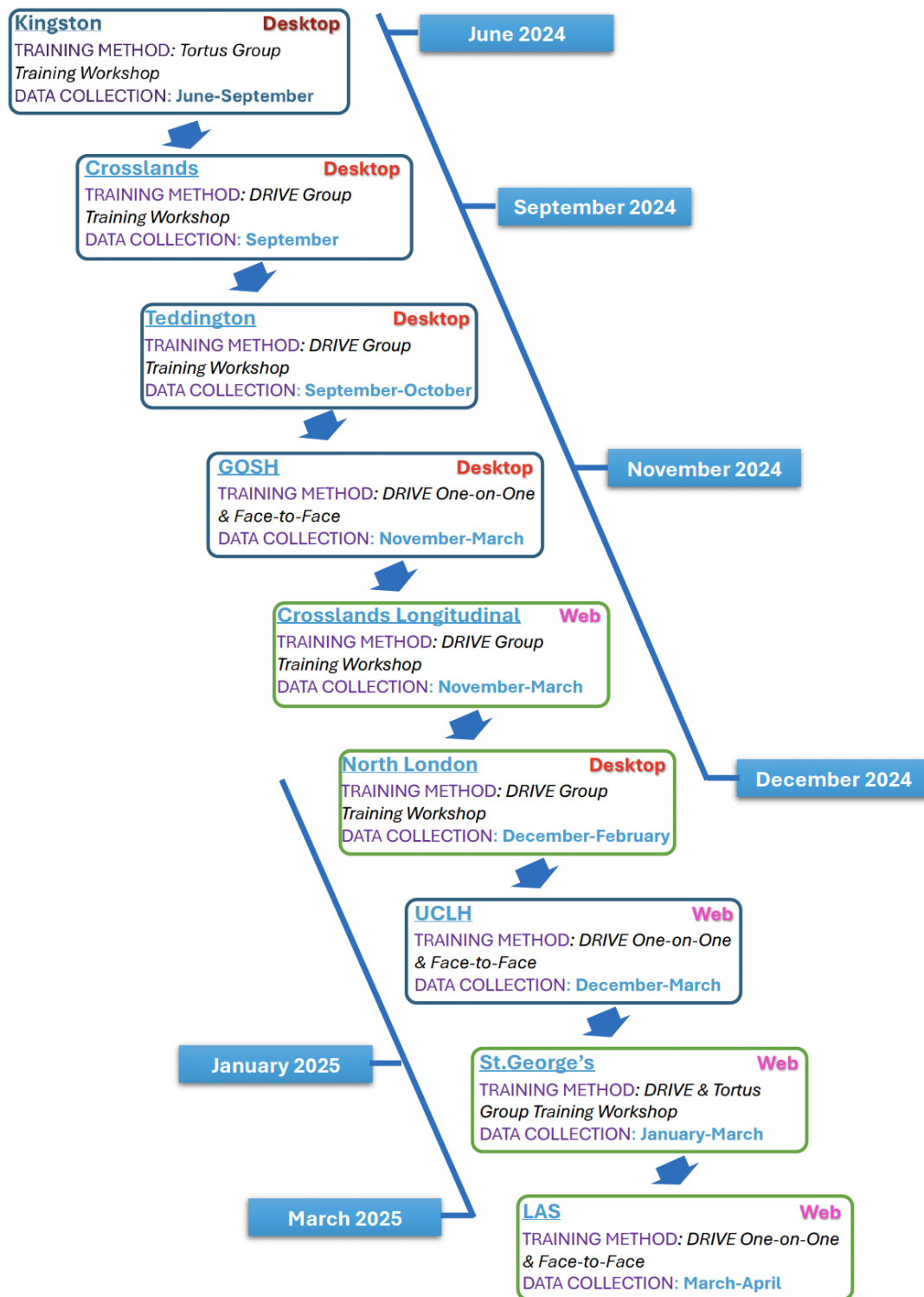


Figure 3.3: Evolution of the study in terms of training methods and software versions

Participation at core and non-core sites

Numbers of clinicians, specialties and episodes of care in both the baseline and AVT conditions for core and non-core sites are shown in Table 3.2. Participation rates among patients were generally high – Table 3.3 provides information for each core site about number of recruited participants, number who were not recruited and reasons for non-participation, if known. A number of appointments across all core settings were virtual, thereby precluding participation for those patients. Participation rate data were not collected at non-core sites. Consent was not formally taken at non-core sites although the use of the AVT technology was explained to participants and they had the option to decline its use during their consultation.

Table 3.2: Sites, number of participating clinicians and patient encounters (any non-core site clinicians who completed 10 or more AVT encounters are included in AVT usage numbers)

Core site – TimeCat observations	Number of participating clinicians (per protocol)	Baseline encounters Number recruited/number invited (%)	AVT encounters Number recruited/number invited (%)
Site 1	19	198/212 (94%)	190/ 214 (89%)
Site 2	7	70 /72 (97%)	70/ 74 (95%)
Site 3*	10	100/ 122 (82%)	103/ 116 (89%)
Site 4	38	381/402 (95%)	381/ 411(93%)
Site 5	30	302/ 305 (99%)	300/ 316 (95%)
Total core	104	1051/ 1113 (94%)	1044/1131 (92%)
Non-core site	Number of clinicians using AVT	Number of AVT encounters	
Site 6	24	4664 AVT encounters	
Site 7	57	3821 AVT encounters	
Site 8	7	172 AVT encounters	
Site 9*	10	7727 AVT encounters	
Total non-core	98 clinicians	16384	
TOTAL AVT encounters	192 clinicians across 9 sites	17428 patient AVT encounters across all sites	

Data are not available from non-core sites for patients/families who declined to participate or where it was deemed inappropriate

*same site included in both core and non-core

Reasons for not consenting to participate included concerns about technology, data security concerns, did not want observer in room, changed mind during consultation and no reason or other reason given.

Table 3.3: Reasons for patients/parents/carers not consenting to participate in the Core study at baseline and AVT

Reason for Not Consenting	Baseline					Baseline Total	AVT					AVT Total
	Site 1	Site 2	Site 3	Site 4	Site 5		Site 1	Site 2	Site 3	Site 4	Site 5	
Concerns about Technology	1		2	1		4	5	1	1	9	8	24
Data security concern						0	5			1		6
Did not want observer in room	6		6	4	1	17	9		2	4	2	17
Changed mind during consultation	6					6						0
No Reason Given / Other Reason	4	2	14	16	2	38	8	3	10	15	6	42
Grand Total	17	2	22	21	3	65	27	4	13	29	16	89

3. Quantitative data

Introduction

This chapter details the analysis of time-motion observations collected via TimeCaT for clinicians at five core sites across London. As stated, this data was collected in order to evaluate two hypotheses: (i) the use of ambient voice technology in a consultation will increase the percentage of direct care delivered during a consultation (vs indirect care), and (ii) the use of ambient voice technology in a consultation will decrease the total time required to see a patient. The analysis focuses on two main variables, total time of session and direct care percentage, compared between Baseline and AVT arms.

Dataset

The dataset consists of time-motion observations collected via TimeCaT for clinicians at five core sites across London. Clinicians were observed across two arms: Baseline (normal practice) and Ambient Voice Technology (practice utilising the AVT Tool). During these sessions, observers recorded clinician time allocation via task buttons within the TimeCat tool. These tasks represented how session time was used, with a task button for *direct care* (speaking to or examining the patient without multi-tasking) and a set of task buttons for *indirect care* (time spent with attention not fully on the patient). These indirect care tasks included options for any task using the computer or creating handwritten notes. The application then output data featuring a unique id, site, time spent on each task, time since observation session began, time spent on each task and notes added. Data were cleaned to include only valid, real observation sessions from clinicians who were observed under both conditions with a minimum of 9 observations at each arm (a 10% margin was agreed for minimum number of observations, allowing clinicians with at least 9 rather than 10 observations to be included). The core analysis was then performed on a per-protocol subset of these clinicians filtering out clinicians who did not reach the required observation count or failed to meet the protocol guidelines. The per-protocol dataset consisted of 2095 observations across 104 clinicians. This time-motion data were joined with data from Monday.com, recording further observation details; observation type (Baseline or AVT), specialty, site, clinician name, observer, appointment type (first vs follow up), translator status (present or not), and accompanied status, which was agreed as people present in the room outside of observer, clinical team and patient, with an allowance for one chaperone for paediatric appointments. These data were fully captured for 4 of the 5 core sites, however use of Monday.com began after Site 1 data collection.

The two variables of interest were total time of session and direct care percentage. The aim of the analysis was to compare these metrics between Baseline and AVT arms.

Total Time of Session

There were two ways of calculating the total session time. The first method used the column '*Duration_Seconds*' which was recorded directly by TimeCat as the time from the observer opening the 'New Observation' page, to clicking 'Finish'. The second method was to use the sum of all tasks recorded during a session. In this analysis we have chosen the latter method for several reasons to most accurately represent the length of the session. Many observers reported that due to lags in the TimeCat system, they had been starting the tool early (pre-session) and then beginning to record tasks when the patient entered. There was also a brief period of protocol change where 'post-patient' time was recorded, as a way of looking at clinician admin time, which would skew the session length if the first method was to be used.

Direct Care Percentage

Direct care percentage was calculated by summing the duration of 'direct care' tasks and dividing this by the total duration of all tasks per observation. Note that for the portion of observations where post-patient time was recorded, there was a cut off that filtered out any tasks started after this time.

Methodology

Time-motion data were sourced from an excel file exported from the TimeCat website, with primary use of the *Obs_index* sheet (for session summaries), *Consultation* sheet (for task breakdowns) and *TotalTime* sheet (for information on post-patient time). Session detail data were sourced from CSVs exported from Monday.com. Data on sites for clinicians was sourced from a CSV compiled within GOSH.

Data Cleaning

- TimeCat data, Monday.com data and the clinician site CSVs were loaded into Python using the pandas and glob libraries.
- Observations incorrectly labelled as test/training data were corrected.
- Training/test sites were dropped from the data, and void clinicians (i.e., study drop-outs, protocol deviators) removed.
- The site column was parsed using regular expressions to extract observation arm, Clinician Name, Department, and Institution.
- Data were cleaned to correct whitespace/punctuation issues, and misspellings.

- The notes column was parsed to identify appointment type, translator status, and accompanied status. Note this was mainly used at Kingston site and was later superseded by the Monday.com boards.
- Observation Data were updated by Monday.com data, which was used as a singular source of truth, so any incorrectly labelled sites/observation types were corrected here.
- Certain task times were manually edited based on notes/feedback from observers after errors.
- Consultation data were merged with TotalTime data in order to add flags to tasks started after post-patient time began, and these were then filtered out.
- Direct care percentage was calculated by summing direct care tasks over sum of all tasks per observation.
- Consultation data were merged with Obs_index data to bring in direct care percentage and summed task total time.

Statistical Methods

Generalised Linear Mixed Models (GLMMs): For our direct care variable of interest, several GLMMs were compiled and compared using information criteria, Akaike information criterion (AIC), and Bayesian information criterion (BIC). The best model was then chosen based on these methods, alongside the coefficient of determination (R-squared) and residual diagnostics.

Paired T-Test: For the direct care variable of interest, we used a parametric paired t-test applied to the per-clinician means at each arm to assess if the true mean difference was equal to zero. A p value of <0.05 was considered statistically significant.

One Sample T-Test: For the total time variable, we used a parametric one sample t-test applied to the relative percentage change in total time, calculated from each clinician's mean session durations before and after the introduction of the AVT Tool, to assess if the average change across clinicians was significantly different to zero. A p value of < 0.05 was considered statistically significant.

Wilcoxon Signed-Rank Test: This non-parametric test was applied to the per-clinician means at each arm to assess if there was a statistical difference in direct care between arms. We also applied this to the relative mean percentage change in total time based on mean values. As this test does not assume normality, this was used for further robustness to consolidate parametric testing.

Global Medians: Global medians were calculated across all observations within each arm group for the dataset of N=104 clinicians. Medians were chosen as the measure of central tendency as opposed to means due to the presence of outliers within the dataset, evidenced in Figure 4.1 and Figure 4.2 in the summary statistics section for direct care percentage and total time below.

Exploratory Analysis

Exploratory analysis was conducted in order to explore distribution of the data and determine the appropriateness of chosen statistical methods. The distribution of contributing variables was analysed across arms to investigate the potential for confounding variables influencing the outcome. It was concluded that the distribution was similar across arms for each contributing variable (i.e. for appointment type; there were similar proportions of first versus follow up appointments at Baseline and AVT arms). This provided greater confidence that the differences observed were due to the effect of AVT.

Random effects were also analysed to investigate the variability between levels of direct care and total time of sessions between sites and clinicians. This informed decisions about the structure of models during the statistical analysis stage. This analysis confirmed that variation was present, with baseline and AVT levels of direct care and total time differing between them. This was important to consider within analysis because the effect of AVT would differ across these levels.

Finally, we investigated the distribution of the dependent variables, to identify skew and shape, so that the validity of chosen tests could be assumed.

Further detail on exploratory analysis can be found in the [Appendix E](#).

Statistical Analysis and Results

Direct Care Percentage

Summary Statistics and Global Medians

The below boxplot (Figure 4.1) comparing direct care percentage between the baseline and AVT arms clearly shows a shift towards higher levels of direct care given after the introduction of the AVT tool. The median direct care increased from 70.0% at baseline to 86.5% at AVT, showing a 16.5 percentage point absolute increase (equating to a 23.6% relative increase).

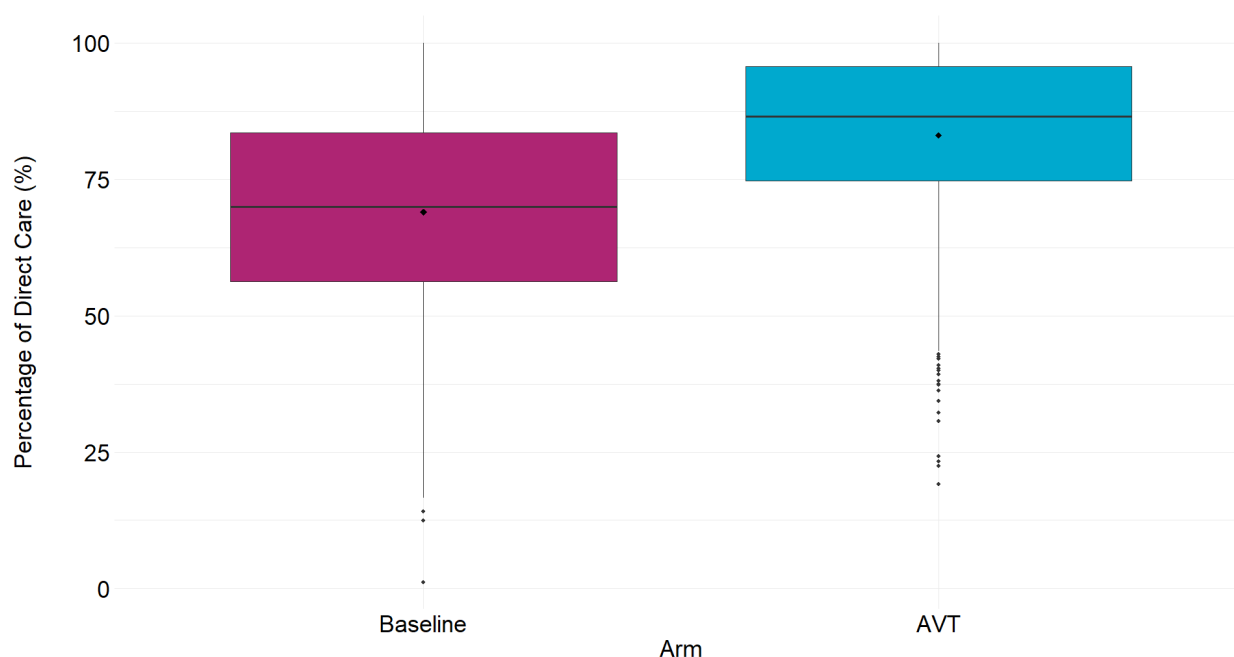


Figure 4.1: Direct care percentage compared across Baseline and AVT arms for all observations in per-protocol dataset

Generalised Linear Mixed Modelling (GLMM)

Initial Considerations

As evidenced in the exploratory graphs, (Appendix E: Figures [E.4](#), [E.5](#), [E.6](#), [E.7](#)), there is clear variation across sites and clinicians. As clinician and site are nested (each clinician only practices at one site), there will be collinearity present here. If both variables were to be included in the model, this collinearity would cause model instability and give skewed estimates. To confirm collinearity, Cramer's V was calculated, which returned the result of 1 – perfect collinearity. Therefore, the conclusion is that only one of these variables would be included in the model. Clinician name was chosen due to its higher granularity, and the fact that it encompasses site-level variation.

Model Exploration

Several model designs were explored across a variety of distributions with fixed and random effects. We selected the final model based on superior fit, indicated by the lowest AIC and BIC, along with satisfactory residual diagnostics. The chosen model was a beta model with *fixed effects*; AVT presence (dependent), appointment type, accompanied status, translator status and total time, and *random effects*; clinician

name. The choice of this beta model was supported by our exploratory analysis of our direct care percentage variable (Appendix E: [Figure E.8](#) and [Figure E.9](#)). As the dependent variable was a percentage and within the 0-100 boundaries, a GLMM with a beta distribution was an appropriate choice. Further details on the [reasoning](#) behind the chosen model and effect structure can be found in the [Appendix E](#).

Model Results

The final beta model found that the use of AVT was significantly associated with a higher direct care time ($p < 0.001$), strongly supporting the hypothesis that AVT influences percentage of direct care.

The model estimated the predicted proportion of time spent on direct care with and without the use of AVT. At baseline, this was predicted to be 69.3%, increasing to 84.1% at the AVT arm. This corresponded to a 14.8+ percentage point change, with a 95% confidence interval of +13.6 to +15.7 percentage points.

Model Residual/Assumption Check

The model fit and validity were evaluated through use of residual diagnostics and plots, which gave the conclusion that the model fit was suitable for this dataset. Further details on residual and assumption checking can be found in Appendix E.

Paired T-Test

A paired t-test was performed, comparing per-clinician means at Baseline and at AVT to assess if the mean difference between conditions was significantly different from zero.

The test produced a highly significant p-value < 0.001 , providing strong evidence against the null hypothesis ($t = 10.72$). This estimated that there was a mean increase of 14.06 percentage points, with a 95% confidence interval ranging from 11.46 to 16.67 percentage points. These results provide strong evidence that the use of AVT was associated with a meaningful increase in direct care percentage.

The assumptions of the t-test were evaluated through use of Shapiro-Wilk and distribution plots, which gave a conclusion that the use of a t-test was indeed appropriate for this dataset. Further details for this testing are evidenced in the Appendix E.

Wilcoxon Signed-Rank Test

To further consolidate the results from our beta model and t-test, a non-parametric Wilcoxon signed-rank test was conducted to assess whether there was a significant difference in direct care percentage between baseline and AVT arms. This test does

not assume normality; therefore, it is suitable given the deviations from normality observed within the data. This gave a highly statistically significant p-value of <0.001, providing strong evidence that the difference in direct care percentage at baseline and AVT was significant.

Total Time

Summary Statistics and Global Medians

The below boxplot (Figure 4.2) comparing total session time between the baseline and AVT arms shows a decrease in total session time (in minutes) after the introduction of the AVT tool. The median total time decreased from 18.4 minutes at baseline to 16.9 minutes at AVT, showing an 8.15% relative decrease.

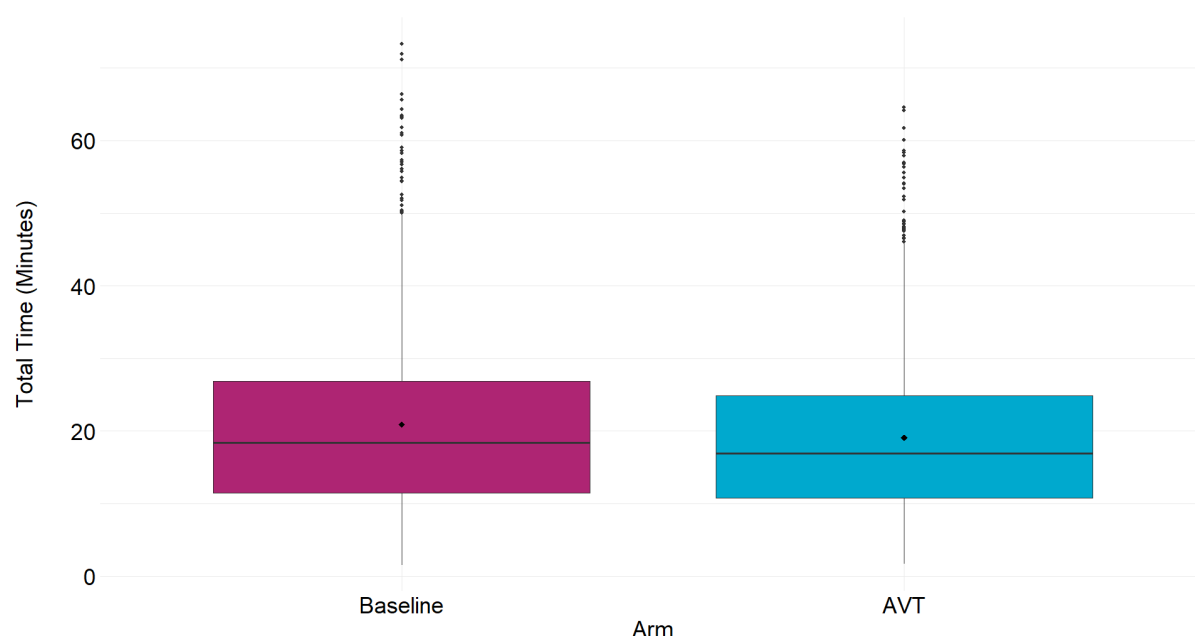


Figure 4.2: Total session time compared across Baseline and AVT arms for all observations in per-protocol dataset

One Sample T-Test

To determine if there was a difference between total session time at baseline and the AVT arm, we performed a one sample t-test on per-clinician mean percentage differences $((B \text{ Mean} - AVT \text{ Mean}) / (B \text{ Mean}) * 100$ for each clinician) to assess if the mean difference between conditions was significantly different from zero.

This test produced a significant p-value <0.005 , which gave us strong evidence against the null hypothesis and indicated that total session time was reduced after the introduction of AVT ($t=3.072$). This estimated that there was an average relative reduction of 5.86%, with a 95% confidence interval of reductions of 2.08% to 9.64%.

The assumptions of the t-test were evaluated through use of Shapiro-Wilk and distribution plots, which gave a conclusion that the use of a t-test was indeed appropriate for this dataset. Further details for this testing are evidenced in the Appendix E.

Wilcoxon Signed-Rank Test

To further consolidate the results from our one sample t-test, a non-parametric Wilcoxon signed-rank test was conducted to test whether the percentage change in session time across clinicians differed from zero. This test does not assume normality; therefore, it is suitable given the deviations from normality observed within the data. This gave a highly statistically significant p-value of <0.001 , providing strong evidence that the difference in total time at baseline and AVT was significant.

Limitations

Limitations applying to the quantitative data can be found in [Appendix E](#).

4. Survey data

Methods

Patient and parent experience measures (core sites only)

Anonymous patient and parent/carer reported experience measures (PREMs) were designed to collect quantitative and qualitative data from patients and/or parents/carers about their experience of the consultation with the clinician. Patient and parent/carer PREMs were collected at the core sites only. The patient and parent/carer surveys consisted of demographic questions, Likert-type questions and a box for free text comments. Numbers of questions differed slightly between respondent groups. Topics covered included time spent by the clinician with them/their child, the use of the computer during the consultation, their ability to ask questions and how well the questions were answered, and how their experience compared with previous clinic consultations at that hospital (unless it was their first appointment). The PREM had good internal reliability (Patient PREM: Cronbach alpha = .777; parent/carer PREM: Cronbach alpha = .807). Patients and parents/carers were also asked to complete the Net Promoter Score (14), indicating their willingness to recommend the use of the AVT technology to their friends and family. Separate PREMs were developed for children and young people (not reported due to low numbers).

Clinician experience measures (core and non-core sites)

Clinician experience measures were similarly designed for both the baseline and AVT elements at both core and non-core sites and included demographic and Likert-type questions and a free-text box for additional comments. Clinicians were asked to provide their name, professional group and specialty and three words that encapsulated their experience of being in the clinic that day. Topics covered by the Likert questions included time spent with the patient, computer tasks during the consultation and satisfaction with the accuracy, completeness and relevance of their clinic notes and letters. Internal reliability was excellent (Cronbach alpha = .917). The NASA Task Load Index (TLX) (15) was also included. The AVT PREMs had additional questions specifically about clinicians' experience of using the AVT together with the Net Promoter Score as an indication of their willingness to recommend the use of the AVT technology to friends and colleagues.

PREMs were analysed using descriptive statistics and baseline and AVT questionnaires were compared using non-parametric statistical tests (all data were non-normally distributed) for either paired (where clinicians had completed both a

baseline and AVT PREM) or unpaired (parent/carer PREMs) data. Scores on the individual subscales of the NASA TLX were compared between baseline and AVT conditions using Wilcoxon signed rank tests. A total TLX score was computed by summing the subscales without weighting and dividing by the number of subscales (Raw-TLX) (16). Free text comments were thematically analysed (17).

Results

Patient and parent/carer experience measures

Table 5.1 provides the number of patient and parent/carer surveys completed at baseline and in the AVT arm of the study, together with the response rate for each site. Figures showing these data are provided in [Appendix F](#).

Table 5.1: Numbers of patient and parent/carer surveys completed at each site; percentages refer to the proportion of participants who completed a survey (response rate)

Core Site	Baseline surveys			AVT surveys		
	Patient surveys (n)	Parent / carer surveys (n)	Response Rate per site (%)	Patient surveys (n)	Parent / carer surveys (n)	Response Rate per site (%)
Kingston Hospital	140	51	71	91	28	56
Teddington Community Care	35	0	50	16	0	23
Crosslands Surgery	45	1	46	34	8	41
UCLH	31	5	11	45	9	18
GOSH		93	24		55	14
Total	251	150	34	186	100	27

Figures describing patient demographics can be seen in [Appendix F](#). Information on other patient characteristics is shown in Table 5.2.

Patients

The distribution of the patient responses by site differed significantly between the baseline and AVT conditions ($X^2=12.2$; $p=.007$) and a higher proportion of respondents in the baseline condition reported anxiety ($X^2= 5.1$; $p=.024$) (Table 5.2).

Table 5.2: Patient characteristics

	Hearing Impairment (%)	Learning Disability (%)	Visual Impairment (%)	Anxiety (%)	Autism (%)	ADHD (%)
	Patients					
Baseline (n=251)	21 (8)	6 (2)	18 (7)	31 (12)	3 (1)	2 (1)
AVT (n=186)	10 (5)	5 (3)	15 (8)	11 (6)	1 (1)	5 (3)
	Parents / Carers					
Baseline (n=150)	9 (6)	18 (2)	11 (7)	7 (5)	16 (11)	7 (5)
AVT (n=100)	5 (5)	11 (11)	3 (3)	5 (5)	11 (11)	3 (3)

Overall, patient responses were very positive across all sites at both baseline and in the AVT condition, with no significant difference in overall satisfaction with the consultation between the two conditions (Figures 5.1, 5.2 and 5.3).

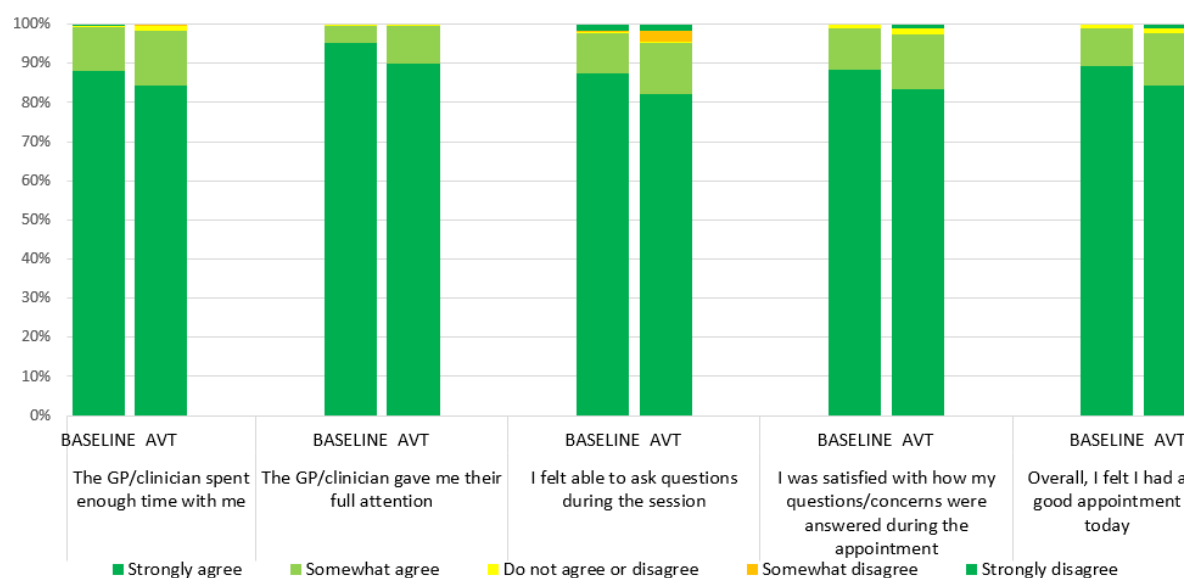


Figure 5.1: Patient experience at baseline and AVT

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings Across the NHS

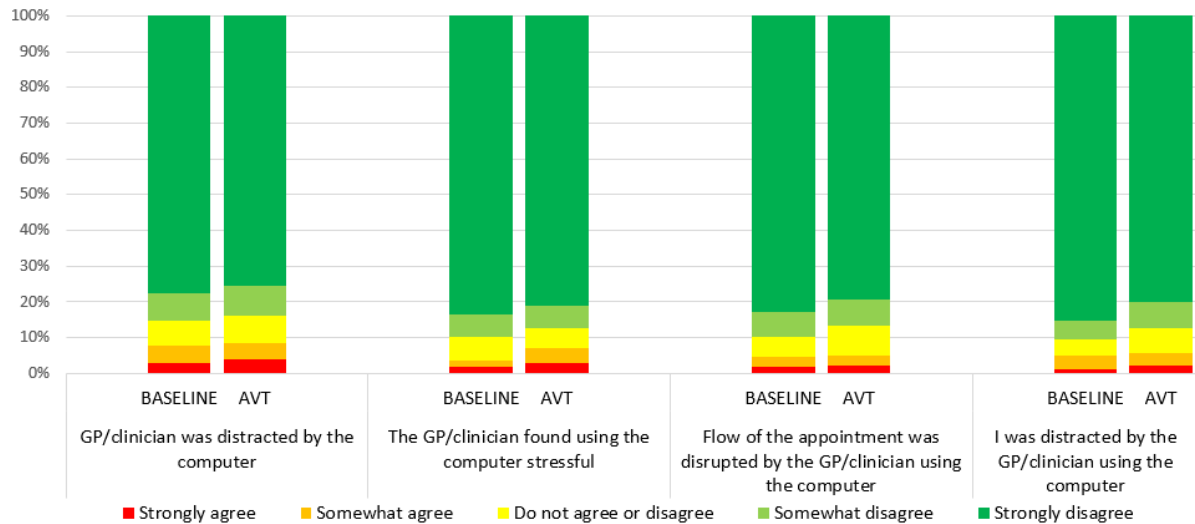


Figure 5.2: Patient experience at baseline and AVT

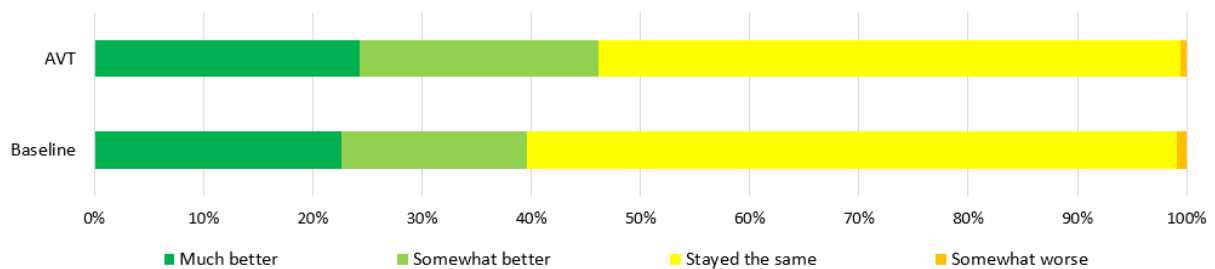


Figure 5.3: Patients comparing current and previous appointment experience: baseline and AVT.

Participants for whom this was a first appointment are not included.

Net Promoter Score - Patients

Patients were likely to recommend the use of the AVT in a GP or clinic visit to a family member or friend, with a high proportion of promoters (n=107; 58%) and relatively few detractors (n=46; 19%) and an overall score of 38.38 (Figure 5.4).

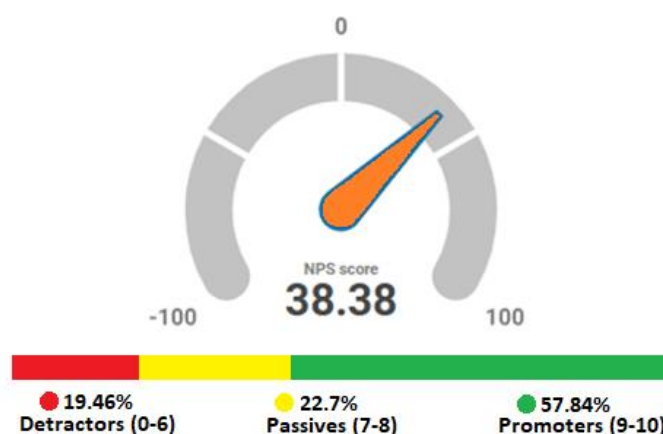


Figure 5.4: Patient Net Promoter Score ([Appendix H](#))

Free-text comments - Patients

Twenty-one patients provided brief comments, which were generally positive about the appointment in general or the use of the AVT. Some patients were not aware of the AVT or did not notice any difference - *"I didn't notice the AI being used – I had a very positive consultation experience"* whilst a few commented on the benefits of the AVT in terms of attention received, *"It was nice for the doctor to give their full attention to me and not tapping away at the computer."* One commented that it was *"Much better than expected, not intrusive"* and the potential of the AVT for the future was also recognised, *"The AI did not interfere at all and if it makes note taking a thing of the past, great"*. Only one patient commented on any changes in their own behaviour as a result of the AVT, enabling them to stay more focused *"I made fewer irrelevant asides know it was recorded"*.

Parents/carers

On the parent/carer survey there were no differences between the baseline and AVT conditions on any patient demographics or characteristics (Table 5.2 and Figures F.4 to F.6, and Table F.2 in [Appendix F](#)).

Parent/carer survey responses were similarly positive across all sites at both baseline and in the AVT condition, with no significant differences between the two

conditions (Figures 5.5 and 5.6). Relative to a previous appointment, experiences were more positive with AVT (Figure 5.7).

Figures describing parent-carer demographics can be seen in Appendix F.

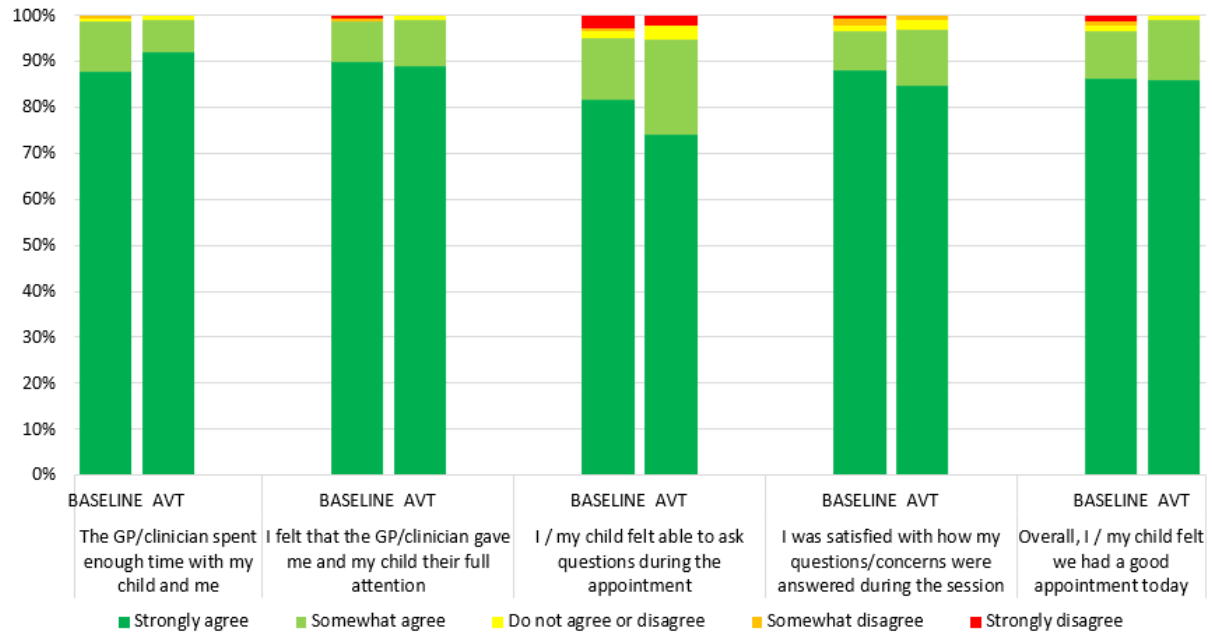


Figure 5.5: Parent/carer experience at baseline and AVT

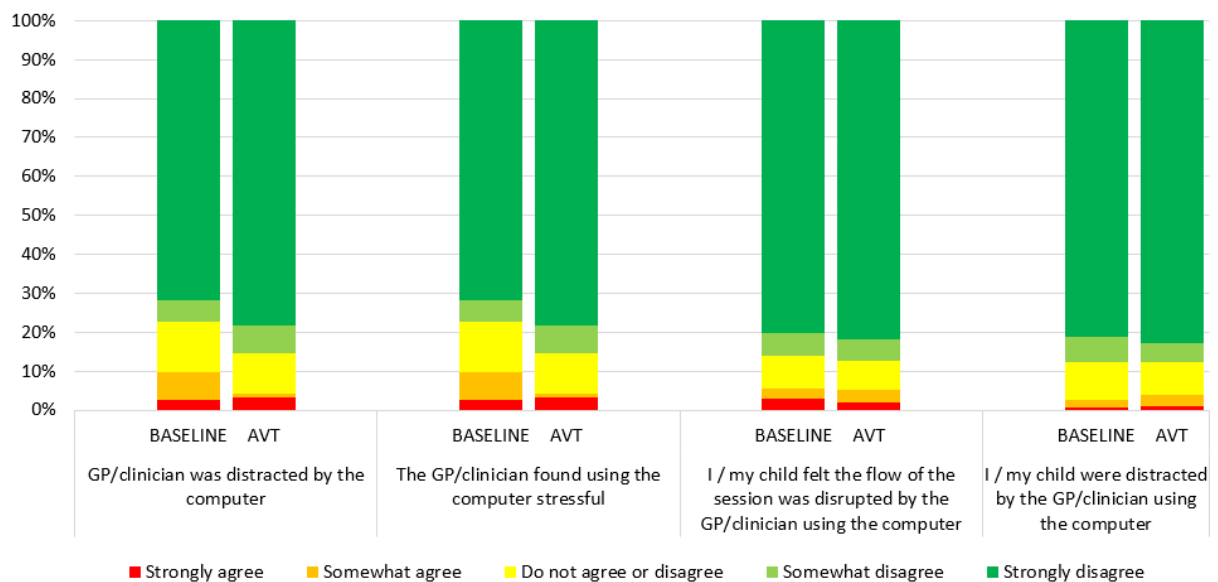


Figure 5.6: Parent/carer experience at baseline and AVT

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings Across the NHS

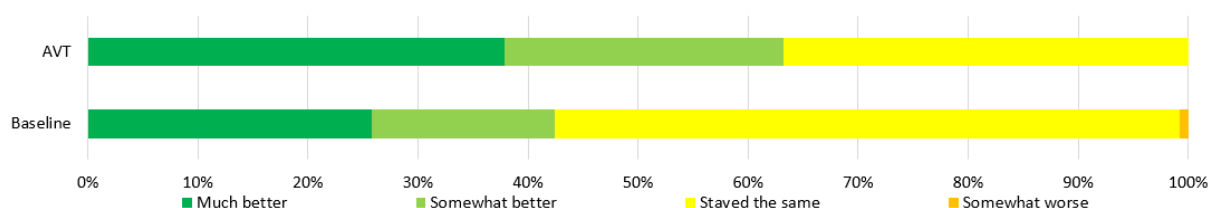


Figure 5.7: Parent/carer comparing current and previous appointment experience: baseline and AVT

Participants for whom this was a first appointment are not included.

Net Promoter – Parents/carers

Parents/carers were likely to recommend the use of the AVT in a GP or clinic visit to a family member or friend, with 64 (64%) promoters and an overall score of 47.47 (Figure 5.8).

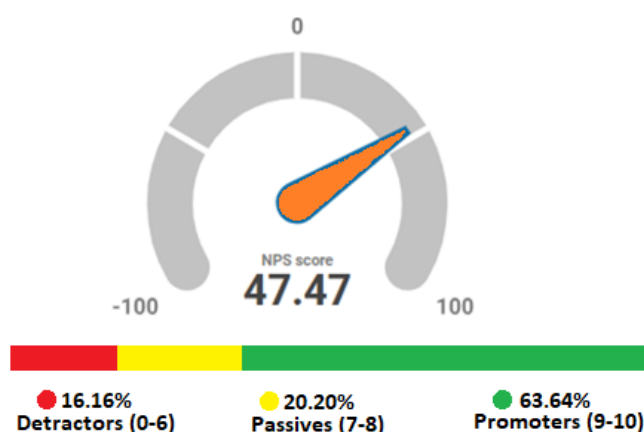


Figure 5.8 : Parent/carer Net Promoter Score ([Appendix H](#)).

Free-text comments – Parents-carers

Twelve parent/carers provided comments on the use of the AVT. Similar to the adult patients, parents commented on improved interactions with the clinicians, *“Just had quality time with the consultant instead of a dictaphone and him typing”*, improved efficiency, *“Very good appointment today and I like the new approach. It is easier to have a speedy appointment due to my child’s disability”* and not noticing the AVT. Some parents also requested greater transparency about what happens to the data, *“It would be good to be able to access detailed information on how data is processed, by whom, and where but overall this is a great initiative”* but overall experiences were very positive.

Clinician experience measures

Description of sample

Two hundred and twenty three clinicians completed a baseline and/or AVT survey, with 122 clinicians completing both baseline and AVT questionnaires across all (core and non-core) sites. Table 5.3 and figures F7 and F8 in [Appendix F](#) provide the number of clinician surveys completed at baseline and in the AVT arm of the study for the core and non-core sites, together with the response rate for each site and an indication of how many clinicians completed both a baseline and AVT PREM.

Table 5.3: Total number of surveys completed at baseline and AVT stage.

	Number of participating clinicians at project initiation	Baseline Surveys Completed (n, %)	AVT Surveys Completed (n, %)	Both Baseline & AVT Surveys Completed (n, %)	Included in TimeCat Per Protocol Baseline + AVT Observations (%)
Core Sites					
Site 1	28	24 (89)	17 (61)	15 (54)	19 (68)
Site 2	7	7 (100)	7 (100)	7 (100)	7 (100)
Site 3	10	10 (100)	10 (100)	9 (90)	10 (100)
Site 4	40	29 (73)	27 (68)	25 (63)	38 (95)
Site 5	35	22 (63)	19 (54)	14 (40)	30 (86)
Total	120	92	80	70	104
Non-core sites					
Site 6	28	26 (93)	19 (68)	17 (61)	N/A
Site 7	50	37 (74)	30 (60)	29 (58)	N/A
Site 8	13	10 (77)	6 (46)	6 (46)	N/A
Site 9	10	N/A**	6 (60)	N/A	N/A
Total	101	73	61	52	N/A

** Site 9 did not complete baseline surveys as they had already done so as a core site

Information about the 32 professional groups and/or specialities of the clinicians who completed a PREM is shown in Table 5.4.

Table 5.4: Clinical specialities and professional groups represented by participants

Clinical Specialities (n=32)	Total
Ambulance Service	84
Audiology	3
Audiovestibular Medicine	1
Breast Surgery	2
Cardiology	2
Clinical Genetics	4
Dental	3
Dermatology	3
Diabetes	2
Dietetics - Paediatrics	2
Emergency Department (ED)	28
Endocrinology	5
ENT	6
Gastroenterology	1
Geriatric medicine	1
GP Practice Consultation	12
Gynaecology	8
Haematology	3
Immunology	1
Mental Health	10
Neurology	7
Neurosurgery	2
Oncology	1
Ophthalmology	5
Orthopaedics	2
General Paediatrics	3
Pharmacology	2
Physiotherapy	10
Plastics	1
Respiratory	5
Rheumatology	5
Specialist Neonatal and Paediatric Surgery (SNAPS)	2
Grand Total	227

Findings

Bar charts are used to show distribution of responses at baseline and in the AVT condition for all surveys completed. Statistics are presented for the paired data (i.e. for the 122 clinicians who completed both surveys).

Clinician experience of using the AVT

The majority of respondents had used the AVT at least once before using it in the clinical situation, with 24 (20%) describing themselves as a first time user of AVT ([Figure F.9, Appendix F](#)).

Experience of using the AVT was generally positive, with >80% agreeing that they felt confident to use the AVT, it worked as expected and it was helpful to use it ([Figure 5.11](#)). Although 30% strongly agreed that the AVT met or exceeded their expectations, 18% disagreed that this was the case. However, the vast majority agreed that the AVT was easy to use ([Figure 5.13](#)).

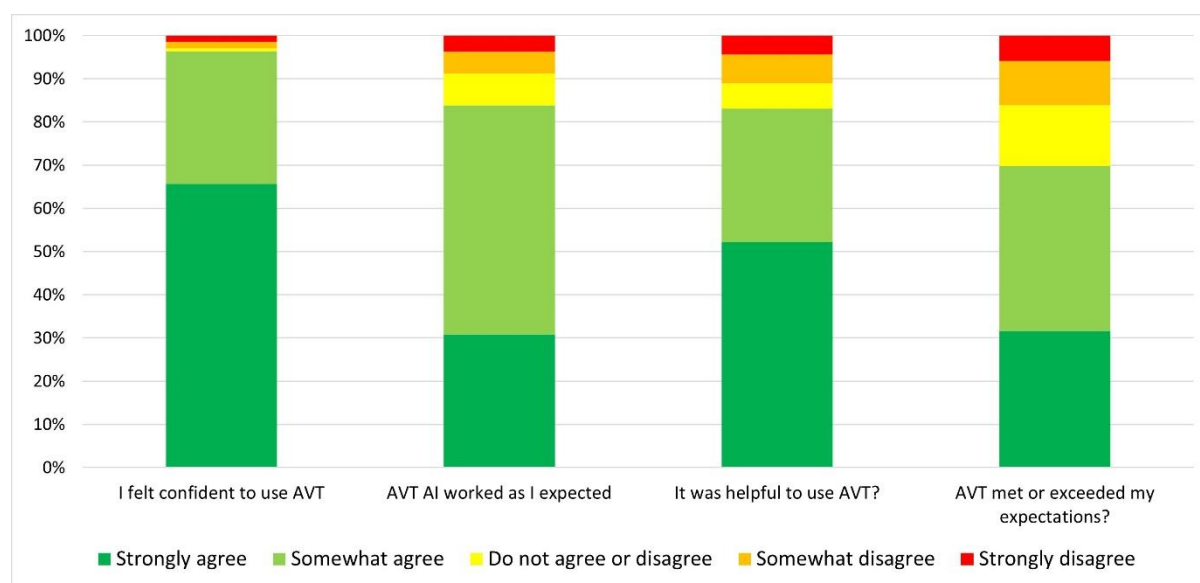


Figure 5.11: Clinician experience of using AVT (n=137, all AVT surveys)

Clinician experience with using the AVT with patients

For the 122 clinicians who completed surveys at baseline and in the AVT arm, there were significant positive changes in the level of agreement that they had sufficient time with each patient ($Z=-4.711$; $p<.001$), they were able to give patients their full attention ($Z=-6.541$; $p<.001$), their satisfaction with the care given ($Z=-4.393$; $p<.001$)

and their overall experience was positive ($Z=-4.524$; $p<.001$) in the AVT compared with baseline conditions (Figure 5.12).

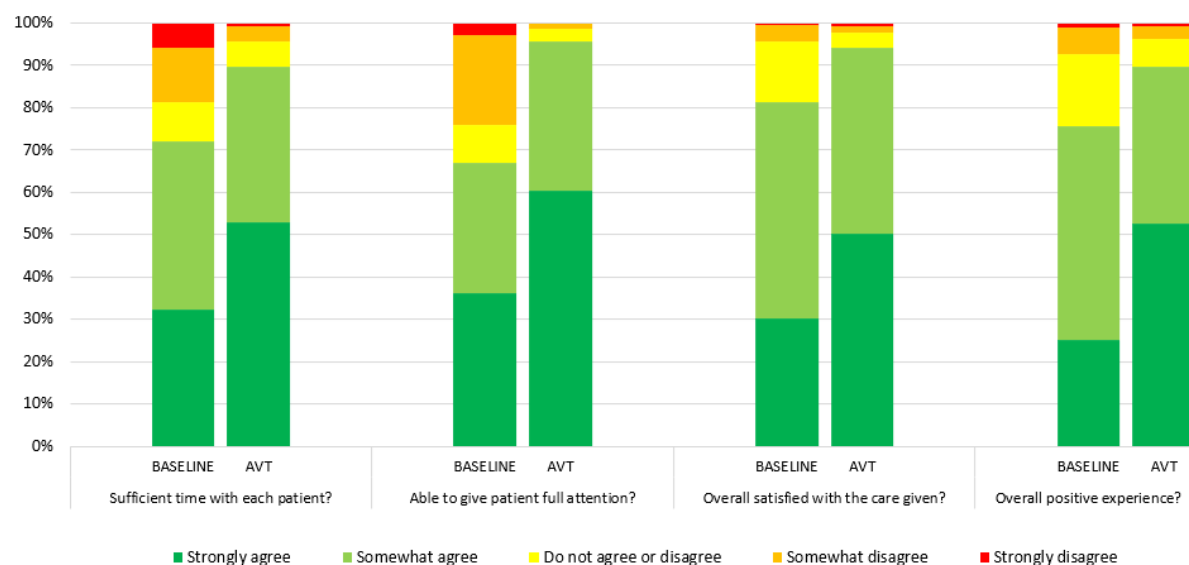


Figure 5.12: Clinician experience with patients at baseline (n=208) and with AVT (n=137)

Clinicians were asked about using the computer during clinic and for all three aspects mentioned, responses were more positive for AVT (distracting: $Z=-5.581$; $p<.001$; stressful: $Z=-6.195$; $p<.001$; disrupted flow ($Z=-6.743$; $p<.001$) (Figure 5.13).

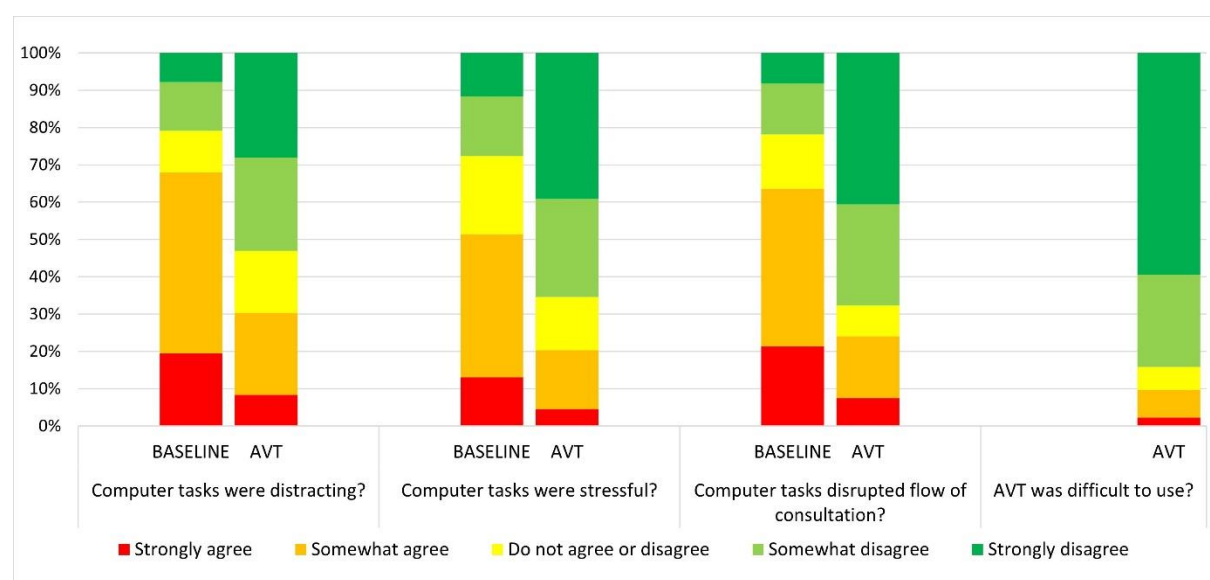


Figure 5.13: Clinician experience in clinic at baseline (n=208) and with AVT (n=137)

Clinical notes

There were no significant differences between baseline and AVT in terms of satisfaction with the accuracy, completeness or relevance of the clinical notes (Figure 5.14). However, satisfaction with the effort of checking the notes ($Z=-6.538$; $p<.001$), time to review/edit the notes ($Z=-7.169$; $p<.001$) and the template of the notes ($Z=-2.248$; $p=.025$) was higher with AVT than at baseline.

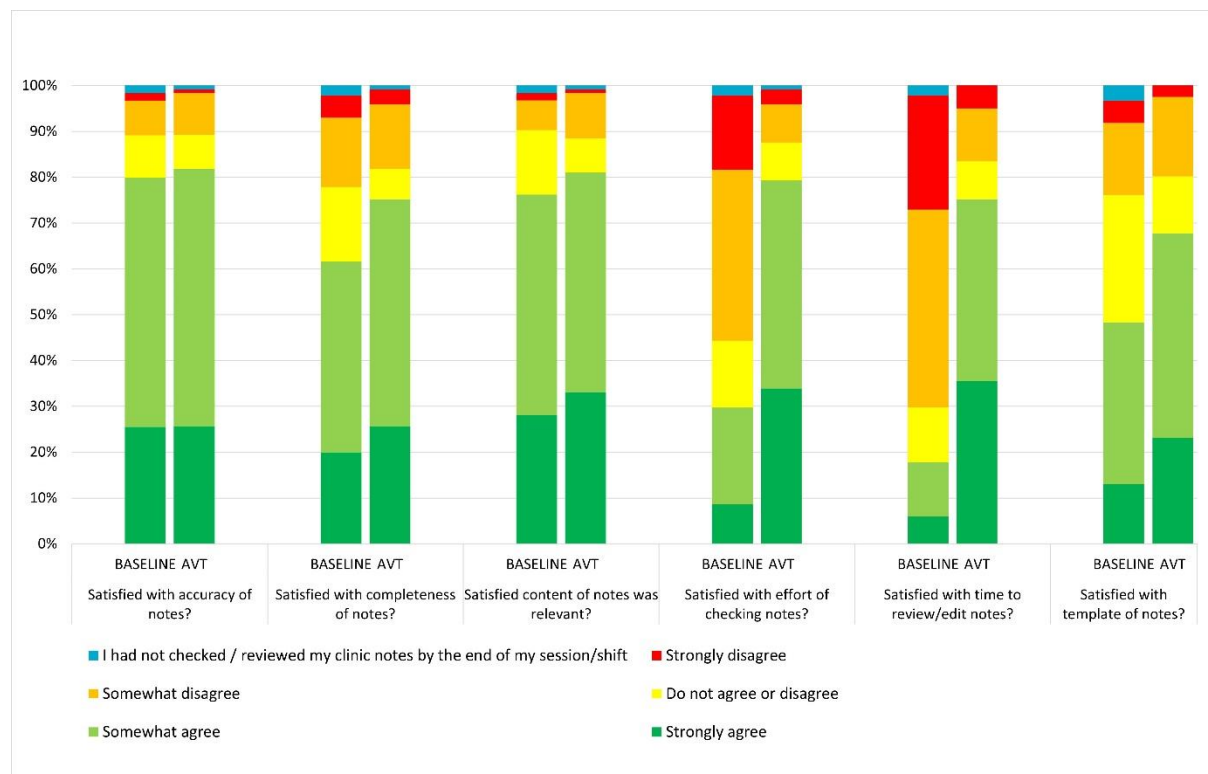


Figure 5.14: Clinician satisfaction with clinical notes at baseline (n=208) and with AVT (n=137) (accuracy, completeness, relevance, effort, time and template)

A higher percentage of the patient notes was completed at the end of the session with AVT (Figure 5.15).

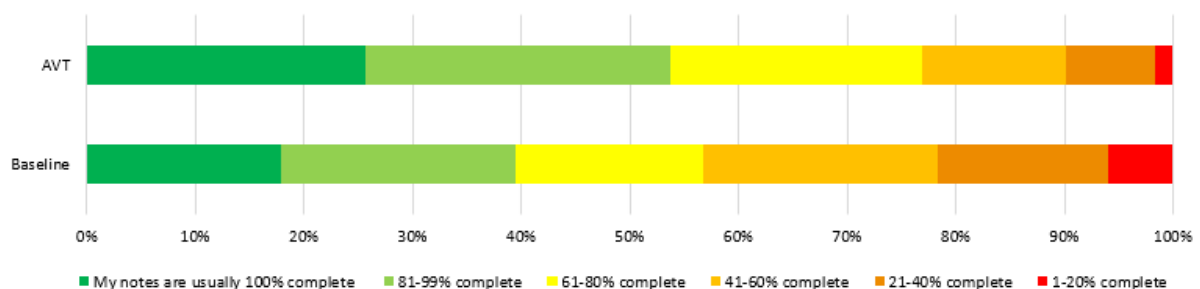


Figure 5.15: Percentage of notes completed by the end of a clinical session at baseline (n=208) and with AVT (n=137)

Clinic letters

A similar, although reduced, pattern of satisfaction with the clinic letters was seen as with the clinic notes (Figure 5.16). Satisfaction with the accuracy, completeness and relevance of the letters did not differ between baseline and AVT but there was greater satisfaction with checking the letters ($Z=-2.643$; $p=.008$) and time to review/edit the letters ($Z=-2.410$; $p=.016$). There was no significant difference between baseline and AVT in terms of satisfaction with the letter template.

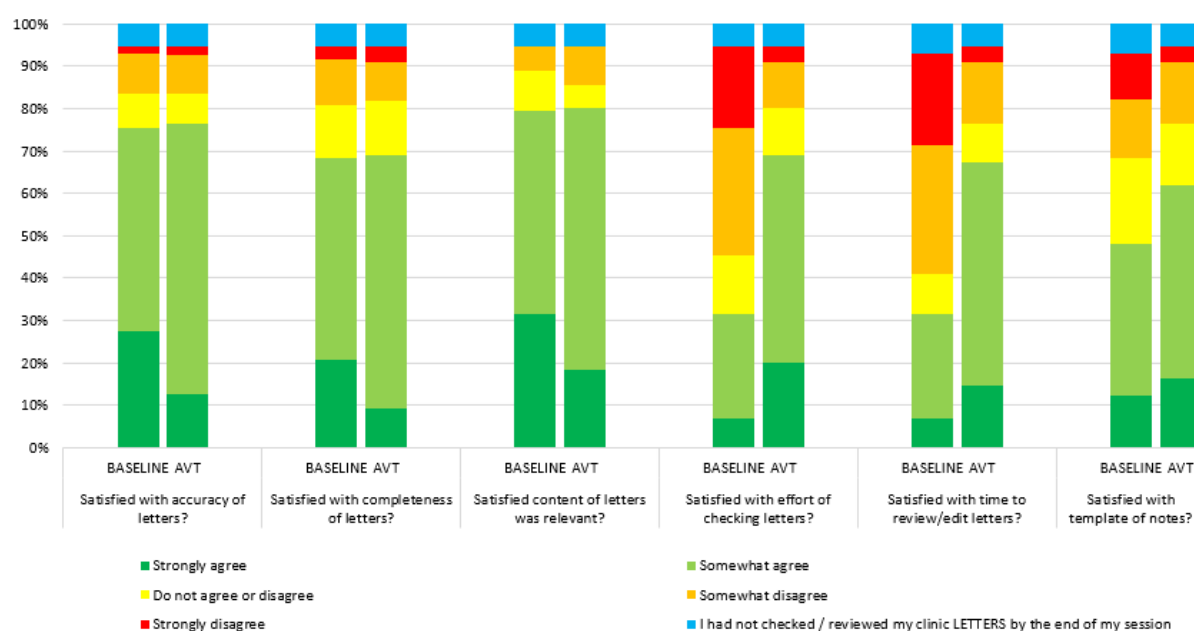


Figure 5.16: Clinician satisfaction with clinic letters at baseline (n=73) and with AVT (n=55)

In contrast to the clinic notes, there was little difference between baseline and AVT in the % of letters which were completed by the end of the clinical session (Figure 5.17).

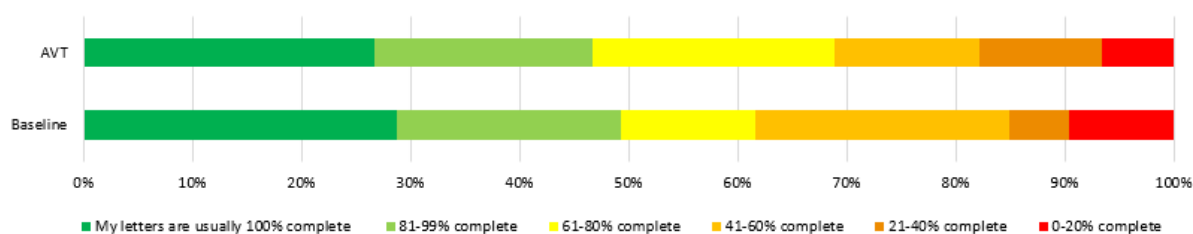


Figure 5.17: Percentage of letters completed by the end of a clinical session at baseline (n=73) and with AVT (n=55)

Clinicians at two sites were asked how overwhelmed they felt by note taking and record keeping (Figure 5.18), with a higher proportion indicating that they did not feel overwhelmed with AVT compared with baseline.

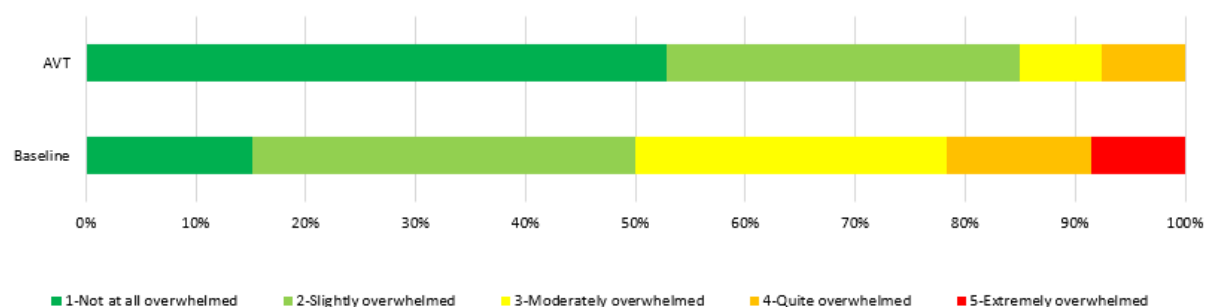


Figure 5.18: Percentage of clinicians feeling overwhelmed by note taking or record keeping at baseline (n=106) and AVT (n=53) (Sites 6 and 7 only)

Net Promoter Score - Clinicians

Clinicians were likely to recommend AVT to a friend/colleague (Figure 5.19), with 43 (41%) promoters, 37 (35%) passives and 25 (24%) detractors. Of note, one of the earliest sites had a high (86%) proportion of detractors, which skewed the overall results and is likely to reflect, at least in part, more issues with training, templates and hardware.

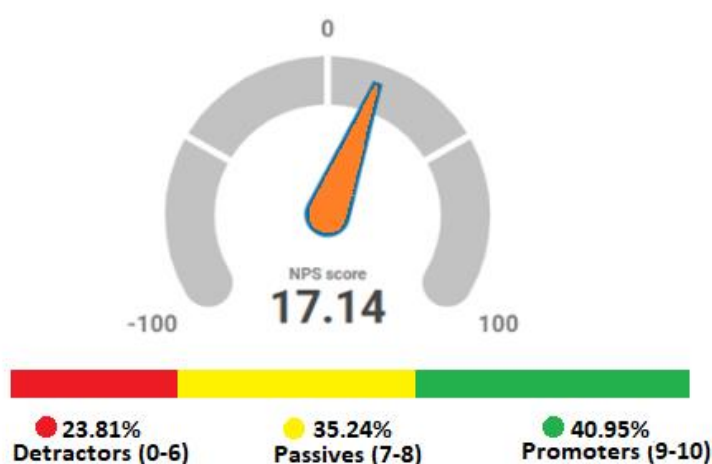


Figure 5.19: Clinician Net Promoter Score

NASA Task Load Index

There was a significant reduction in total NASA cognitive load score for the paired sample in the AVT condition compared with baseline (baseline median: 5.83; IQR: 1.79; AVT median: 5.08; IQR: 2.00; $Z=-5.398$; $p<.001$). There were significant reductions in 5 of the 6 subscales – performance was not significantly different. Figure 5.20 provides NASA subscale and raw TLX scores for the unpaired (total) sample.

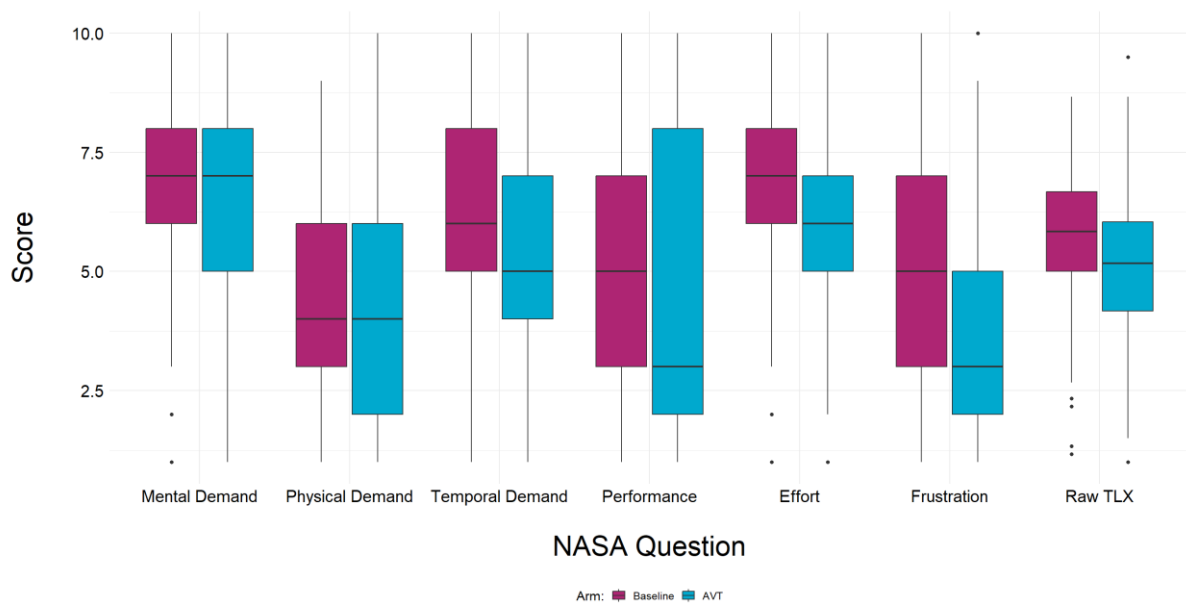


Figure 5.20: NASA subscale and raw TLX scores for total sample

Free-text comments - Clinicians

Ninety-three clinicians provided free text comments and comments were received from at least 5 clinicians from each site. Thematic analysis resulted in five themes and eight sub-themes (Figure 5.21), with both positive and negative views expressed. Some themes were consistent across sites, others were more nuanced and specific to particular sites. Below is a brief summary of each theme with illustrative quotes.

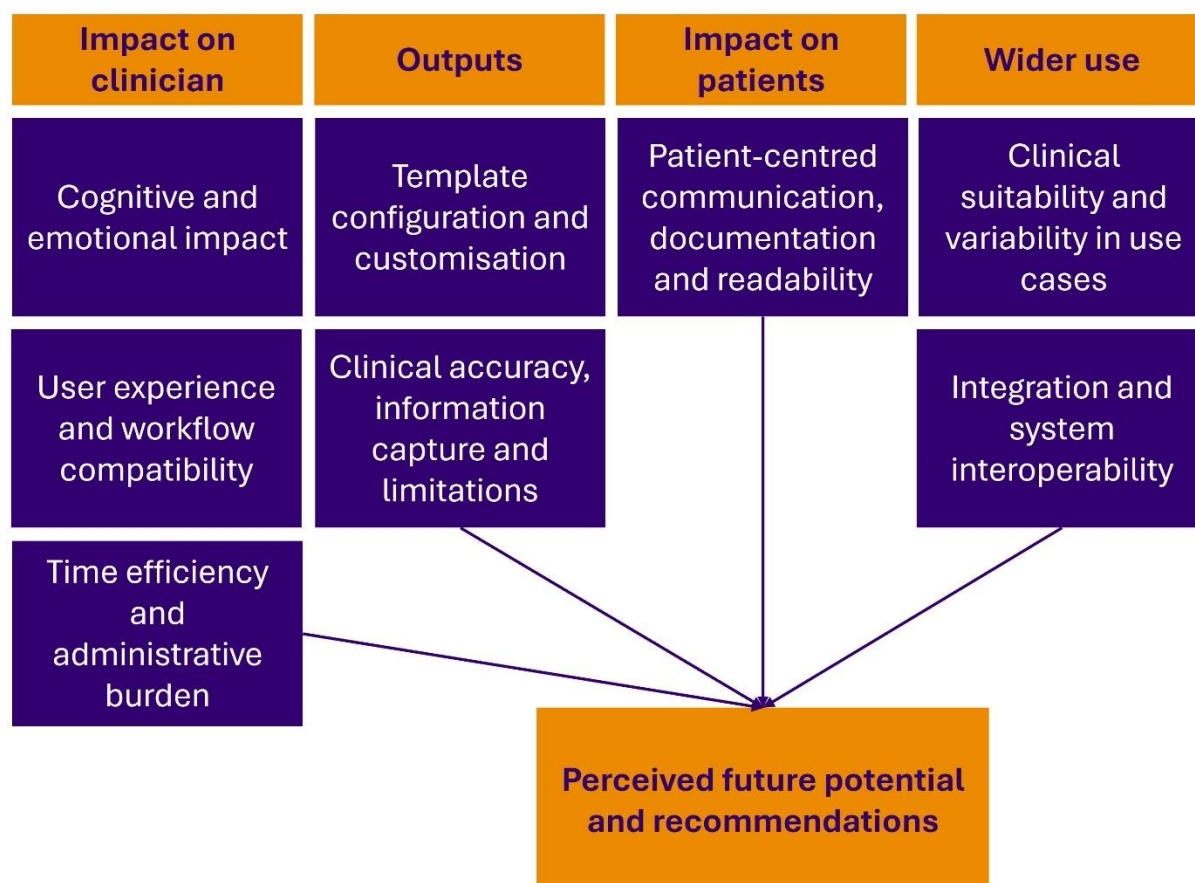


Figure 5.21: Themes from free text comments

Theme: Impact on clinician

Cognitive and emotional impact

- Clinicians from all settings reported reduced cognitive load, increased focus during consultations, and improved emotional wellbeing.

“I love the [AVT]. I feel more confident the more I use it. I feel it takes some stress away from the additional cognitive strain of capturing everything I said. It completes notes with a high level of accuracy and safety netting that is suited to me style. I think it has benefits for the mental aspect for the clinician and improved my general feeling of seeing several complex patents one after another than can be psychologically draining.” – Site 3

“[AVT] has taken a huge administrative burden away and allowed me to complete notes accurately without struggling to recall. This is especially

important when at peak fatigue. I hope it can be a permanent fixture in the ED.” – Site 6

Neurodivergent clinicians, including those with dyslexia, found the system especially beneficial, citing significantly reduced stress and improved documentation quality.

“[AVT] is the first software I have used as a dyslexic that works. I used to finish my shift with a feeling that my brain had been torn in two, with trying to assess patients and write notes. Now after [AVT] I am coming away from work much calmer in my mind and not needing recovery time to feel normal in my mind again. Now my focus is 100% on my patient, and therefore they are getting a better assessment from me, better care, and my notes are much more comprehensive because I'm not struggling to fight the dyslexia.” — Site 7

- In some cases technical challenges (particularly early on in the trial) led to frustration, increased workload, and emotional strain.

“We did have a few occasions where [AVT] would log itself back out to the Home Screen thus deleting what had already been recorded mid way through a recording.” – Site 7

User experience and workflow compatibility

- Usability varied widely, influenced by environmental factors (e.g., noise), hardware reliability (microphone quality, battery, device compatibility), and clinician familiarity with the tool.

“My issues were not so much the software but the hardware — the mic stopping picking up, cutting out, running out of battery, not working on particular computers. Because then the transcript isn't complete and so translates it inaccurately.” – Site 6

- Some clinicians identified software limitations in managing multiple voices or chaotic scenes involving families and interpreters.

“Struggles with thick accents or when more than two people talking. Very good if patient has ability to discuss in a structured way with important information.” – Site 3

- Many clinicians wanted to be able to pause/resume recording or to view live AI capture mid-consultation to ensure completeness.

“Would be nice to reopen a consultation and continue where you left off instead of it being closed.” – Site 7

Time efficiency and administrative burden

- Comments about time efficiency and administrative burden were mixed. For many, AVT significantly reduced documentation time, enabled real-time note capture, and allowed for more efficient post-consultation administration. Others found that the time spent proofreading, editing, and reformatting notes offset the anticipated efficiency gains, particularly in complex or follow-up consultations.

"[AVT] has drastically reduced the amount of time spent documenting clinic sessions and I am confident I could reduce this further if I create my own templates." — Site 8

- Others noted no time savings due to technical or environmental issues.

"I find it much easier to dictate into the AI after seeing the patient than actually using the AI during a consultation... multiple times probably due to loud environment or not great English from patient or for other factors, when using during consultation the AI would miss a lot of key info and so spent a long time editing its documentation." — Site 6

- Some noted that while initial summaries were helpful, the editing demands for letters offset time benefits.

"I was very happy with the notes made by [AVT] with almost no editing needed. The letters required much more editing I think this could be improved by changing the style used and the information included in the letter template." — Site 3

Theme: Outputs

Template configuration and customisation

- A major focus of the comments was on the template design, which, when it did not meet the needs of the clinicians, was a significant source of frustration across all settings due to the misalignment between clinical reality and AI output.

"I did not have a good template set-up to use [AVT] - I gave it my previous clinic letters as a template but the letters generated by [AVT] missed the key aspects of the consultation, also they were too waffly for the aspects that were correct." — Site 4

- Some clinicians expressed a preference for bulleted formats with NHS abbreviations, highlighting inefficiency when forced to edit full sentences.

“The [AVT] tech works well and it does a good job of capturing the relevant information from the consult and putting it into the notes, but I think that for our team the notes are too wordy and long-form. Our team likes our notes to be very dot-point based with a few words as possible, using lots of abbreviations and [AVT] is the opposite, using full sentences and long form answers for all sections. I was not able to get my template to change enough to match my note preferences.” –

Site 2

- Customisable templates were requested for diverse clinical presentations and patient cohorts. Some clinicians advocated for pre-visit configuration or auto-generated templates based on clinic type.

“Creating new templates not easy in clinic - would be good to have library of templates which we can modify.” – **Site 5**

“Unfortunately the template settings were difficult to make to ensure the notes were formatted in a way that we are so used to.” - **Site 2**

“The way it summarises the data is super useful. The templates are very good. Overall I think it’s fantastic.” – **Site 7**

Clinical accuracy, information capture and limitations

- There were many positive comments about the performance of the AVT in capturing and summarising routine consultations accurately, including with a translator

“The voice capture was perfect (even with a translator on one occasion).” – **Site 4**

- Some clinicians expressed concerns about the AVT’s ability to manage complex cases, capture nuance, and accurately interpret multiple voices, accents, or emotionally charged conversations.

“The use of [AVT] in mental health is limited by the fact that the software is not able to capture the nuances in patients' presentations, such as mental state. It needs dictation by clinician for formulation and risk assessment of cases as the software is not able to do that reliably.”

– Site 8

- Omissions were highlighted in developmental and psychosocial detail when patients were children/young people, use of inappropriate language complexity for patients (especially adolescents), and failure to accurately reflect family dynamics or educational details. Specific clinical details, such as medication names or key symptoms (e.g., vomiting frequency, safeguarding issues), were sometimes omitted or inaccurately represented.

“I have had some issue with accuracy which I suspect is related to the template. In paediatrics it’s fairly important how many times a patient has vomited in the last 24 hours, when they last vomited, the duration of vomiting, how much per day, it seems to get confused and just pluck a number and says they vomited x times.” – Site 6

- Some concerns were raised about the tendency of the AVT to either oversimplify or generate overly verbose outputs, both of which impaired clinical utility.

“The AI itself truncates the conversation down to the point where most of the details are lost.” – Site 7

“The notes are wordier than need to be. It struggles with specific physio outcome measures.” – Site 1

Theme: Impact on patients

Patient-centred communication, documentation and readability

- Clinicians described how AVT enhanced clinician–patient interaction through improved eye contact and reduced screen time.

“After it accurately recorded the first couple of sessions, I stopped making my own documentation and was able to give the patient all my attention which made a significant difference to the quality of the consultation.” – Site 8

- There were repeated concerns across sites about the tone, structure, and readability of AI-generated letters—often described as “too AI-like” or inappropriate for direct patient communication. Clinicians commented that ‘their voice’ was lost in the AI-generated outputs.

“Letters to patient use unnecessarily complex language—particularly for adolescent patients... I have written clinic letters from scratch as this has been quicker than trying to edit [AVT] generated letters.” — Site 6

- Clinicians commented on the need for layperson-friendly summaries for patients.

"It's very helpful and most of the effort is about trying to configure a template that accurately captures my voice. Part of this is providing a better summary to non specialist clinicians and families and this helps me in this regard. I usually write quite technical letters and the AI makes the letter easier to understand." – **Site 4**

Theme: Wider use

Clinical suitability and variability in use cases

- The AVT performed best in structured environments such as GP surgeries or well-controlled paediatric outpatient clinics, particularly with less complex patients.
- It was less reliable in unstructured, dynamic settings (e.g. ambulance scenes, EDs, mental health crisis consultations), where real-time adaptability and nuanced data capture were critical.

"[AVT] excels in calm primary /urgent care presentations where there is time to sit opposite a patient and run through history taking and assessment. Some of the challenges present in complex cases in consultation mode where there maybe multiple relatives or bystanders providing history or talking on behalf of the patient." — **Site 7**

Integration and system interoperability

- Clinicians from all sites expressed a need for better integration with local electronic patient records.

"If the formatting of the [AVT] notes could be reflected in the EPR (EPIC) that we use, this would help greatly and reduce the amount of time taken to format the notes, and generate legible letters." – **Site 6**
- Some clinicians noted workflow disruption due to the web-based nature of the app and its incompatibility with EPRs.

"It would probably be better either integrated into EPCR or as an app, as having it as a webpage in the background would sometimes cause it to log out." – **Site 7**

Theme: Perceived future potential and recommendations

- There was considerable enthusiasm and optimism across settings for further development and broader adoption, contingent upon:
 - o Improved EPR integration
 - o Real-time usability enhancements
 - o Flexible, clinician-driven customisation
 - o Better handling of complex, multi-dimensional data

"I love AI as a tool and would happily use ambient listening and note-taking for every patient. It helps with almost every aspect of patient care... However, this would only be the case if the programme was effective." —

Site 7

5. Interview Data

Methods

Clinicians from each site were invited to participate in individual semi-structured interviews or small group/focus group discussions to discuss their experience of using the AVT once the trial of the AVT technology had been completed at their site. Interviews were either face to face or virtual, dependent on clinician preference and availability. Topic guides were used to inform the interviews and focus groups, with additional questions added for individual sites to reflect any specific ways of working. Interviews were audio-recorded with consent and transcribed for subsequent analysis. Framework analysis(18) was used, comprising the five steps of 1. data familiarisation, 2. framework identification (involving both inductive and deductive approaches), 3. indexing, 4. charting and 5. mapping and interpretation. Two members of the team led the qualitative analysis, with other team members contributing to transcription and sense checking.

Results

Participants

Fifty-five clinicians took part in an individual or group interview, with 39 sessions being held in total. The number of participating clinicians from each site ranged from 3 to 9.

Themes

Themes and subthemes are presented below Figure 6.1. Four groups of factors were perceived to have a direct impact on the utility and quality of the output and on the clinician, with utility and quality also impacting the clinician and the patient/family.

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings
Across the NHS

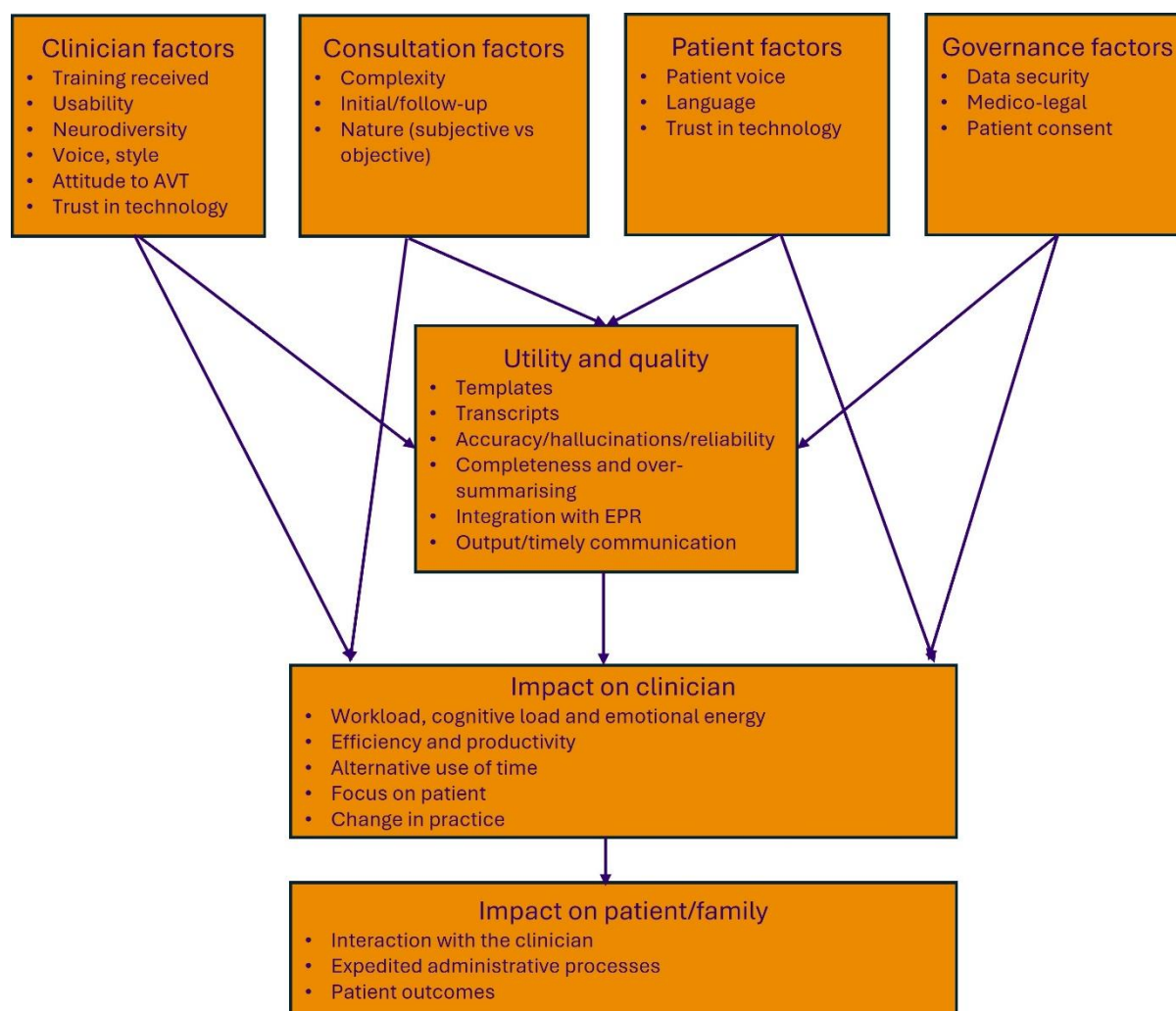


Figure 6.1: Themes and subthemes from the clinician interviews

Table 6.1 provides summary narrative text and illustrative quotes for each of the themes and subthemes, starting with the factors that influenced perceptions of utility and quality and the resulting impact on the clinician and patient/family. Whilst some of the factor subthemes highlight some challenges and requirement for workarounds and modifications to how the AVT was used in the consultation, the impact on clinicians and patients was generally extremely positive. Of note, these data were collected across the 12 months of the trial, during which there was ongoing evolution of the training, template development, software versions and hardware.

Table 6.1: Themes, subthemes and illustrative quotes from the clinician interviews

Theme	Subtheme and explanatory text	Participant quote
Clinician factors	Training received Clinicians trialling the AVT highlighted the importance of sufficient investment in effective, personalised training and timely, practical support in shaping their experiences.	<p><i>“I just signed up and then I sort of felt like I was committed to the whole thing and didn’t know what the time commitment was going to be, didn’t know what the output was going to be... it always just creates a bit of uncertainty” (Site 5_Cln_01)</i></p> <p><i>“They [the trainers] were really, really good. They were amazing. They messaged me (with) as much support as I needed” (Site 5_Cln_02)</i></p>
	Usability Clinicians reported a wide range of positive and less positive experiences regarding the usability and practical integration of AVT across diverse clinical contexts, particularly in multi-speaker and high-acuity consultations or where additional detail needed to be captured or other features integrated.	<p><i>“I used it on a major trauma... it captured everything” (Site 7_Cln_04)</i></p> <p><i>“If there’s a husband and wife... I’ve had to ditch the AI” (Site 3_Cln_01)</i></p> <p><i>“One of the few things I miss about paper... is just being able to draw” (Site 6_Cln_08)</i></p> <p><i>“Able to translate it all into English which was really impressive” (Site 5_Cln_05)</i></p>
	Neurodiversity AVT was seen as particularly intuitive and beneficial for those who were neurodivergent,	<p><i>“It evens out some people’s challenges, if they’ve got dyslexia or are neurodiverse” (Site 6_Cln_02)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	offering significant reductions in administrative burden. Clinicians highlighted the capacity of the AVT to standardise notes and support clinicians who struggle with traditional documentation methods.	<i>"With my dyslexia that is really helpful. It means that I'm a lot less stressed" (Site 7_Cln_04)</i>
	Voice, style Clinicians emphasised the importance of maintaining their personal clinical style and voice in the notes and letters generated by the AVT. While many recognised the utility and time-saving potential of AVT, they commented that the outputs did not always align with their individual preferences or established documentation practices – e.g. the tone was too impersonal or generic, formatting was not how they wanted. However, the potential of the AVT was recognized if it could be adapted to better reflect individual documentation styles.	<i>"It felt like I was reading a letter from someone else's clinic" (Site 1_Cln_01).</i> <i>"It doesn't sound like me. It doesn't write with me" (Site 4_Cln_03)</i> <i>"If it allowed me to... say 'I want a bullet point list'... it would be able to have my personality in my notes" (Site 3_Cln_01)</i> <i>"The AI..... created a nice letter, I like the way it flowed, it took that thinking away. But there were small details that didn't fit how we usually write letters, with a clear plan" (Site 2_Cln_02).</i> <i>"What's being generated is good enough but... it just sounds like it's written by AI." – Site 8_Cln_01</i>
	Attitude to AVT	<i>"I think it's the future. I have no doubt that within five years this will</i>

Theme	Subtheme and explanatory text	Participant quote
	Clinicians expressed predominantly positive attitudes towards AVT, frequently describing it as transformative, promising, and well-aligned with the future of clinical documentation. For some, practical constraints and concerns regarding readiness were seen as current barriers to adoption.	<p><i>be exactly how clinics are recorded and documented” (Site 5_Cln_04)</i></p> <p><i>“A brilliant idea... it really is brilliant” (Site 2_Cln_02)</i></p> <p><i>“I’d like to think there is a future... but there would have to be a lot of improvements” (Site 2_Cln_01)</i></p>
	<p>Trust in technology</p> <p>Clinicians were cautious in trusting the AVT but trust increased with use. For some, confidence grew considerably but others had reservations due to errors or omissions. Issues with hardware, early software glitches and connection issues, and concerns about losing data impacted trust.</p>	<p><i>“Did I think it would work? No, not at all... But very pleasantly surprised and shocked” (Site 7_Cln_01)</i></p> <p><i>“By the end of the second clinic... I probably stopped [taking manual notes] because I was actually so comfortable that it was capturing... all of the conversation and summarising it well” (Site 5_Cln_03)</i></p> <p><i>“it said mother rather than father... that taught me to be extra careful” (Site 4_Cln_07)</i></p>
Consultation factors	<p>Complexity</p> <p>The complexity of the consultation influenced AVT’s effectiveness. Complex cases involving more than one diagnosis revealed limitations in the AVT’s ability to capture and</p>	<p><i>“I think the ambient AI would be much better for somebody like surgeons who don’t need to go into the nitty-gritty, whereas it’s different for a rheumatologist or a neurologist or an endocrinologist who asks about all the systems. Whereas we</i></p>

Theme	Subtheme and explanatory text	Participant quote
	structure diverse clinical content.	<p><i>have so many minutiae to deal with.” — Site 1_Cln_04</i></p> <p><i>“When there are multiple problems ... especially when there are emotional impacts ... I’ve had to really specify that ... the ambient AI.” — Site_3_Cln_05</i></p> <p><i>“I had one clinic where everything was fairly straightforward, and it was brilliant I went home and I said oh this is amazing, this is really good.Then I had a much more complex clinic and I found it was actually, positively unhelpful because it had missed things” Site_4_Cln_07</i></p> <p><i>“So much of medicine is fairly complicated and multifaceted ... Would I trust [AVT] to listen to my consultation, save it and me not be able to edit it? Not at all. Would I rather [AVT] do a summary that sometimes needs a bit of tweaking, sometimes is perfect, sometimes is terrible, that I edit — over me manually writing everything? Yes, every single day of the week!” — Site 6, Cln_01</i></p>
	<p>Initial/follow-up</p> <p>Clinicians generally found AVT to be more effective during initial consultations, particularly in capturing</p>	<p><i>“The ambient AI was really useful for new patient assessments where we do a much longer subjective history... It was really accurate with</i></p>

Theme	Subtheme and explanatory text	Participant quote
	<p>subjective histories, symptoms and patient narratives. AVT performed well in transcribing these conversations accurately and meaningfully. In contrast, its utility during follow-up consultations was more variable. Follow-up appointments were described as quicker and more focused, often involving brief checks rather than detailed narratives. Sometimes the verbosity and workflow of the AVT was perceived as more of a hindrance.</p>	<p><i>recording that for us... for getting the story" (Site 1_Cln_06)</i></p> <p><i>"I thought for follow-ups it was excellent because for follow-up it's a two-minute job..." (Site 1_Cln_04)</i></p> <p><i>"For the follow-ups, not that helpful, because there isn't much of a story and you want that as a brief summary rather than a whole big story." (Site 1_Cln_05)</i></p> <p><i>"From a [AHP] point of view it struggled a little bit more... we would quite often formulate that as bullet points... instead the [AVT]reported everything as quite wordy sentences." (Site 1_Cln_06)</i></p>
	<p>Nature (subjective vs objective)</p> <p>Clinicians consistently reported that AVT was particularly effective in capturing subjective data which are spoken aloud during the consultation. In contrast, objective data such as physical examination findings, measurements or visual assessments were often either omitted or inaccurately recorded by the AVT. The necessity of verbally describing objective assessments was also</p>	<p><i>"It definitely made the subjective faster... it was nicer to be able to communicate a little bit more directly with the patient." (Site 2_Cln_99)</i></p> <p><i>"The patient comes in here, depending on how long their story is, but typically you spend a good half of the session going through signs, symptoms, medical history, drug history, social history, all that kind of stuff. That's a bit that the AAI picks up really well." Site 1_Cln_05</i></p> <p><i>"It was missing aspects of the objective findings that we don't always talk about... it would pick up</i></p>

Theme	Subtheme and explanatory text	Participant quote
	introduced as a workaround to ensure they were captured but for some this needed further refinement.	<i>other bits that were less clinically relevant.” (Site 1_Cln_05)</i> <i>“You are treating very robotically saying out loud ‘now I’m doing left single straight leg raise’...” (Site 2_Cln_99).</i>
Patient factors	Patient voice Clinicians commented on the value of AVT to capture patients’ exact words, lending authenticity to clinical notes. However, there were concerns in contexts where the emotional depth or linguistic complexity of patient narratives might not be adequately captured.	<i>“It was actually very good at recording and picking things up and quite true to how the patient would say it... that’s exactly how they said it as opposed to how I would paraphrase it in my mind”(Site 1_Cln_05)</i> <i>“It will take that first-person view, the patient said, in inverted commas, you know, ‘I had chest pain this morning when I woke up at 6am’, it will pick those points, and it will do a nice story. I do prefer that, because you get more of the patient’s words about why they’ve called” (Site 7_Cln_06).</i> <i>“Most of the stories we hear ... are traumatic... in order to do a ... story justice... you write the story in their words, and you don’t leave out anything. There’s a lot of detail... [AVT] would summarise it in two sentences”(Site 8_Cln_01)</i>
	Language	<i>“I had a family where both parents had learning difficulties ... the transcript is terrible because I think</i>

Theme	Subtheme and explanatory text	Participant quote
	<p>The AVT's transcription accuracy was poorer with patients for whom English was not a first language or those where speech was fragmented or included slang or swear words. Some clinicians introduced work-arounds which were very effective.</p>	<p><i>the syntax that they were using ... was just not what it [AVT] was used to...and the notes it generated were minimal" (Site 4_Cln_07).</i></p> <p><i>"I have started summarising to the patient when the patient speaks poor English or English with a lot of slang. I'm saying to them 'okay, so what you really want to talk about today is the headache, it's got a lot worse', and I wouldn't normally have done that..." (Site 3_Cln_01).</i></p>
	<p>Trust in technology</p> <p>Clinicians reported generally high levels of patient and family trust and acceptance. Most clinicians described minimal resistance to the technology, with only a small number of patients declining participation. Clear communication, transparency about data handling, and respect for individual concerns were highlighted as important for fostering trust.</p>	<p><i>"It doesn't record, doesn't keep a permanent recording of it, are you happy?... I've not had any pushback" (Site 7_Cln_06)</i></p> <p><i>"You might get certain groups of patients... that are sceptical over the technology and they might be worried about use of data and so on" (Site 5_Cln_05)</i></p>
Governance	<p>Data security</p> <p>Data security was an important theme in clinician reflections. While some expressed initial scepticism,</p>	<p><i>"I was very sceptical. Like, it's recording me... seeing the information governance side of it—seeing that it wipes the iPad when you log out, it doesn't save anything</i></p>

Theme	Subtheme and explanatory text	Participant quote
	<p>most reported growing confidence in the system's security features after understanding how data are handled – particularly the temporary use followed by automatic deletion of the data - and the ability to explain that to patients. Clear information governance practices were highlighted as being essential for building both clinician and patient trust in the technology.</p>	<p><i>and the prompt saying nothing gets uploaded, that helped” (Site 7_Cln_04)</i></p> <p><i>“The fact that you could actually say to a patient, all this information will be gone in 24 hours was actually very good... It is literally just a scribe tool for me... then deleted. Actually, I think [that's] reassuring to patients” (Site 5_Cln_06)</i></p>
	<p>Medico-legal</p> <p>Most clinicians viewed the technology as a valuable tool for enhancing accountability, documentation accuracy, and legal defensibility, particularly in complex or high-risk clinical environments. The benefits to both clinicians and patients were highlighted, together with recognition that the medico-legal utility of AVT will likely depend on striking the right balance between automation and clinical responsibility.</p>	<p><i>“If we're going to be honest... transcription stored somewhere on the patient's record... is a protection for the patient and the doctor... I've had a couple of (situations) where they've misinterpreted what I've said” (Site 5_Cln_07)</i></p> <p><i>“Clinicians can find shortcuts just about to do anything... anything that avoids clinicians feeling that they need to take shortcuts [and] copy-and-paste is really positive for me” (Site 7_Cln_03)</i></p> <p><i>“Will it affect the outcome of my ability to defend myself if I made the incorrect decision? Massively yes... I'm cognitively cleaner. I'm less likely to make a mistake” (Site 6_Cln_05)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	<p>Patient consent</p> <p>Patient/parent consent rates were high for the use of the AVT, although views diverged as to what patients should be told. Several clinicians argued that informing patients about AVT use is both necessary and ethically important. Others saw the technology as no different from existing digital documentation systems and therefore felt explicit disclosure was unnecessary.</p>	<p><i>“...most people were happy and consented. We didn’t have any refusals” (Site 5_Cln_07)</i></p> <p><i>“I think the patients should be informed that the conversation is being recorded... they need to know” (Site 1_Cln_03)</i></p> <p><i>“We don’t ask them when they came in about which clinical systems can be used... I don’t see any difference” (Site 3_Cln_01)</i></p>
Utility and quality	<p>Transcripts</p> <p>Transcription accuracy was consistently highlighted and described positively in terms of recording subjective histories and medication details but the degree to which the transcript was used varied. While some found transcripts to be a reliable reference, especially for retrieving missed details or clarifying medication regimens, others rarely referred to them—either due</p>	<p><i>“Without the transcript, that [detail] was missing... So in that sense, yes, another benefit would be having something to look back upon.” (Site 5_Cln_02)</i></p> <p><i>“The transcript was so long... it was just easier to remember what happened” (Site 5_Cln_08)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	to time constraints or confidence in the summary.	
	<p>Templates</p> <p>Templates generated considerable discussion with all clinicians, with consistent highlighting of the critical role of well-designed, personalised templates in shaping the usability and clinical relevance of AVT-generated summaries and documents. Several clinicians found that custom templates tailored to their specialty and workflow could significantly improve output. However, creating, refining, and managing these templates proved to be time-consuming and technically challenging for many users. Clinicians often expressed that they lacked the time or technical support to configure templates to meet their requirements, with concerns about the rigidity of templates. In many cases, clinicians described a disconnect between transcript accuracy and how information was translated into the template or note format. Several expressed</p>	<p><i>“Overall, the template is where the work needs to be, and we didn’t really have enough time at the beginning of our trial to understand that.” (Site 2_Cln_Group)</i></p> <p><i>“The transcript mostly was okay, it was the conversion to the template that wasn’t quite right” (Site 5_Cln_02)</i></p> <p><i>“I’d done what I thought were good templates... I rapidly realised that they weren’t as good as I thought.” — Site 4_Cln_03</i></p> <p><i>“Once your template is what you want it to be... it’s been really, really useful.” — Site 5_Cln_06</i></p> <p><i>“We have worked with our template... since we’ve been doing that, it’s got better about understanding what things we would want to be captured.” — Site 7_Cln_05</i></p>

Theme	Subtheme and explanatory text	Participant quote
	confidence in the raw transcript but dissatisfaction with how AVT interpreted and structured this information. Of note, early users of AVT had more challenges with templates but this improved through the trial with increased training and support for template building.	
	<p>Accuracy, hallucinations and reliability</p> <p>Accuracy was poorer in more complex consultations (highlighted above) and hallucinations were a persistent cause of anxiety, albeit a rare occurrence, including fabricated diagnoses, incorrect clinical impressions, and misattribution of symptoms or treatments. A frequent concern was failure to reliably capture “negatives” — information explicitly indicating the absence of symptoms or risk factors. Linked to this, many clinicians emphasized the importance of proof-reading and clinician accountability. Despite these concerns, some clinicians reported</p>	<p><i>“If it hallucinates or it makes mistakes – and it has definitely hallucinated... that’s quite a big deal” (Site 5_Cln_02)</i></p> <p><i>“There is a real risk if it misinterprets what the patient said and turns it into a clinician’s recommendation” (Site 5_Cln_03)</i></p> <p><i>“We like to have exclusions in our notes... no weight loss, no red flags... we found it quite difficult to get the ‘No’s into the list” (Site 2_Cln_Group)</i></p> <p><i>“My fear is that busy clinicians will just cut and paste it over... and miss things” (Site 4_Cln_04)</i></p> <p><i>“It remembered something about her drugs that I had completely forgotten... little intricacies that I had 100% forgotten... and completely converted immediately” (Site 7_Cln_01)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	that AVT often captured useful information that they themselves had forgotten, thereby enhancing their own reliability.	
	<p>Completeness and over-summarising</p> <p>Clinicians reported varied experiences related to completeness and over-summarisation of AVT-generated documents. Some described the benefits of the AVT documenting more information than they would normally include but more frequently clinicians expressed concern about the tendency of AVT to over-summarise and filter out clinically important details. This was particularly evident in consultations requiring a descriptive narrative or where multiple problems were described.</p>	<p><i>"It was really helpful when I had a follow-up, for example, because my notes on a follow-up appointment are quite limited, but the AI tended to capture a lot more of the things that I would normally put into my notes" (Site 5_Cln_07).</i></p> <p><i>"I might do a really lengthy consultation and then what comes up at the end is not enough in the way of detail" (Site 1_Cln_02)</i></p> <p><i>"The [AVT] just doesn't capture any of that really... a lot of the social stuff was seen as excessive to the consultation so it's been missed out" (Site 3_Cln_01)</i></p> <p><i>"The [AVT] wasn't picking up the bits I wanted to, and it would pick up other bits that were less clinically relevant" (Site 1_Cln_05)</i></p>
	<p>Integration with EPR</p> <p>Clinicians frequently expressed concerns about integration with electronic health records and the lack of this in their organisation. There was broad agreement</p>	<p><i>"The [AVT] system doesn't integrate with our electronic healthcare record system, so it has no prior data about the patient" (Site 5_Cln_01)</i></p> <p><i>"The perfect technology... would be if I could say after the consultation, I need to do an order for this, I need a</i></p>

Theme	Subtheme and explanatory text	Participant quote
	that meaningful, smarter and dynamic integration was not only desirable but essential for scaling the use of AVT in practice and for the full potential of AVT to be realised.	<p><i>form for this, and then the software does it for me” (Site 1_Cln_03)</i></p> <p><i>“The use of AI on its own is one thing, [but] if you combine it with Epic, it’s brilliant” (Site 5_Cln_02)</i></p>
Impact on clinician	<p>Workload, cognitive load and emotional energy</p> <p>Most clinicians reported a dramatic reduction in administrative workload, allowing them to complete tasks that would otherwise take hours or days. AVT alleviated the cognitive burden of documentation with clinicians speaking of being freed from the mental juggling of remembering, summarising, and rephrasing clinical details. The reduction in multitasking not only enhanced focus but also reduced the risk of errors and omissions, also reducing cognitive load and stress and increasing emotional energy. Clinicians described how the mental load of documentation previously bled into their personal lives but AVT</p>	<p><i>“I normally take about three weeks to do all my letters from clinic and I was getting them done within a few days. Sometimes by the end of that clinic, by the end of that day... dramatically different” (Site 4_Cln_01)</i></p> <p><i>“A heck of a lot less thinking was required because it was already there... quicker than dictating a letter straight out” (Site 1_Cln_01)</i></p> <p><i>“You’re handing over patients or taking history at peak fatigue... it’s so easy to miss something at 6am when you’re hungry and tired... to not have to worry about that recall is much better” (Site 6_Cln_09).</i></p> <p><i>“I go back from night shifts now and I feel like I can sleep better... I don’t have that classic post-night stress that you’ve missed something” (Site 6_Cln_07)</i></p> <p><i>“It makes a big difference for me because I could be going over this, giving me a headache, making sure</i></p>

Theme	Subtheme and explanatory text	Participant quote
	<p>reduced that. In contrast, there were a few clinicians who had a different experience because editing or reconciling AI generated output with their usual template increased workload.</p>	<p><i>my notes are right. [Without AVT] I'm double checking it constantly... It makes it more enjoyable" (Site 7_Cln_06).</i></p> <p><i>"I actually started finding it was taking me longer because I had the AI summary and I couldn't let go of my Epic template... in the end I was like, have I really made things quicker for myself?" (Site 5_Cln_08)</i></p>
	<p>Efficiency and productivity</p> <p>Improvements in efficiency resulted in improvements in productivity. Clinicians reported being able to see more patients in the same time frame by recouping time and cognitive bandwidth. However, the increased efficiency raised concerns for some clinicians about escalating expectations to see more patients, highlighting the importance of having time for clinical reflection and emotional processing.</p>	<p><i>"I would say it's probably upped my calls by about 30%, not only because I've cut down on the admin time, but mostly because by not having the fatigue... you've got the ability to go on and have more patient interactions" (Site 7_Cln_05).</i></p> <p><i>"I feel like, my goodness, we're at capacity... Are we going to be expected to have patient after patient after patient? You can't just keep going... I need to process it. Part of typing it up is processing the information" (Site 8_Cln_02)</i></p>
	<p>Alternative use of time</p> <p>AVT in clinical environments produced a nuanced impact on clinicians' use of time – from a redistribution of time</p>	<p><i>"It reduced probably my admin by 80%... then I can go, 'Right, I've got a bit of time between patients. I can quickly chase up that email'" (Site 5_Cln_06)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	toward additional patient care or near real-time documentation to being able to finish their working day earlier or avoid completing notes in their personal time.	<p><i>"Although I don't think I've saved time... the NHS Trust is getting more doctor time out of me" (Site 6_Cln_01).</i></p> <p><i>"What was really helpful... I was doing all my notes usually at the end of clinic or the end of the day, so actually it massively shortened my day" (Site 3_Cln_01)</i></p>
	<p>Focus on patient</p> <p>Clinicians reported overwhelmingly positive impacts of AVT on their ability to engage with patients and families during clinical consultations. The technology significantly reduced the need for active typing or note-taking during the consultation, allowing clinicians to maintain more eye contact, observe non-verbal cues, and hold more natural and uninterrupted conversations.</p>	<p><i>"It was excellent to be able just to concentrate on the patient... you can focus on the patient. You can spend more time reading the nonverbal cues of the patient instead of just constantly being there, looking at the computer" (Site 1_Cln_02)</i></p> <p><i>"I think that the primary benefit for me has been about how much time I can spend interacting with the patients and the families" (Site 5_Cln_03)</i></p> <p><i>"I was completely into the conversation and the consultation because I didn't have to type... It improved my quality of work, the quality of the interaction and probably also the quality of the consultation" (Site 4_Cln_02)</i></p>
	<p>Change in practice</p> <p>Using AVT resulted in a number of adaptive changes in practice, such as</p>	<p><i>"My questioning and assessment has been better, because I want not only the patient to understand, but</i></p>

Theme	Subtheme and explanatory text	Participant quote
	<p>increased verbalisation of clinical reasoning and examination findings and being 'more present' with patients, resulting in positive changes to the clinician-patient relationship and improved communication. However, some clinicians initially experienced some discomfort with abandoning traditional note-taking habits whilst others described a need to restructure consultations to ensure essential content was verbalised, such as describing functional assessments or sensitive findings, which they previously would have noted silently.</p>	<p><i>also the AI to understand what I'm saying" (Site 7_Cln_05)</i></p> <p><i>"I didn't have to type while I was seeing the patient... I was more relaxed, and I was completely into the conversation and the consultation" (Site 4_Cln_02)</i></p> <p><i>"I was surprised to find I did not know what to do with my hands, because it felt very unnatural to not write anything down... I lost my flow a couple of times" (Site 1_Cln_04)</i></p> <p><i>"You wouldn't be able to say [concerning features] with the patient there... so after the patient had left... you would just tell [AVT] what to add in" (Site 2_Cln_Group)</i></p>
Impact on patient / family	<p><i>Interaction with clinician</i></p> <p>Benefits to patients of an improved interaction with the clinician were identified. Clinicians described that, as a result of the positive impact of the AVT on them, they were able to engage in more empathic communication with the patient.</p>	<p><i>"These consultations are so precious... if I don't have to concentrate on typing, the effect on me is positive. The effect on the patient is better because I can spend time fully explaining, using gestures, body language, eye contact" (Site 5_Cln_02)</i></p> <p><i>"It [AVT] did enable me, as a therapist, to use my listening skills even further... I think it improved the</i></p>

Theme	Subtheme and explanatory text	Participant quote
		<i>engagement that we had” (Site 5_Cln_06)</i>
	<p>Expedited administrative processes</p> <p>Clinicians described how the AVT could result in improved efficiency and clarity in relaying treatment plans to patients and other clinicians, potentially reducing the interval between a specialist assessment and action by the GP or other clinician and reinforcing patient understanding of their own treatment plan.</p>	<p><i>“It used to take weeks to dictate, get typed, correct and send... now the GP has the letter by the next day in many cases. That definitely means they can start the blood test or prescription sooner” (Site 6_Cln_04)</i></p> <p><i>“If I said it out loud to the patient, it was able to translate that into a letter that was quite understandable... That will be the main advantage – changes in medication plans to be communicated more quickly and efficiently to everyone. Other teams, the GP, and the patient”(Site 4_Cln_01)</i></p>
	<p>Patient outcomes</p> <p>Some clinicians linked AVT indirectly to improved outcomes via perceived improvements in consultation quality. Others also acknowledged that the expedited administrative processes could have an impact on patient outcomes through more timely prescriptions for example. However, AVT was not yet seen as directly impacting</p>	<p><i>“There is definitely good evidence to suggest that (if) patients feel that they are well cared (for), taken care of, it can improve their outcomes. It is an extrapolation, but if you think that the quality of the consultation is better, the patients and the families feel listened to” (Site 4_Cln_02)</i></p> <p><i>“Now that I can do the letter during the consultation or right after, it means it’s off my desk immediately... which can really help especially when you’re changing drugs or asking for investigations” (Site 3_Cln_01)</i></p>

Theme	Subtheme and explanatory text	Participant quote
	measurable clinical outcomes.	<i>"If what you're asking me is 'do I think it will make a difference in terms [of] directly using that tool', no. But indirectly recording outcomes, I guess there probably is a way of leveraging it, but we're quite a long way off knowing what outcomes it is we want to come with" (Site 4_Cln_04)</i>

6. Individual case examples

Emergency Department (ED)

One of the non-core sites was an emergency department (ED) in a large inner-city teaching hospital. Three key performance indicators were identified for the ED department in relation to the AVT trial (Table 7.1)

Table 7.1: Key Performance Indicators for the ED department

Key performance indicator	Desired outcome	Metric
KPI 1 Number of patients seen by clinician	<i>Improved productivity</i>	Log of each patient seen before and after Ambient Voice in the ED.
KPI 2 Time taken to document clinical note	<i>Improved productivity</i>	Log of time taken to complete first clinical entry on EHR, before and after trial.
KPI 3 Time to clinician being assigned to decision made to being clinically proceedable	<i>Markers for improved patient flow: -Average length of stay (LOS) of patient from clinician assigning to 'Clinically ready to proceed'.</i>	Time taken from when clinician assigns themselves to patient to point a decision has been made that the patient is clinically ready to proceed i.e. discharged, bed booked, transport booked.

During the pilot study there were **4664** ED patient encounters with AVT.

ED - Key Performance Indicator 1

There was a statistically significant increase in the number of patients seen per shift, with an average increase per clinician shift from 9.10 to 10.36, a 13.43% improvement ($t=-4.47$; $p<.001$).

The box plot below (Figure 7.1) shows a clear upward shift in the number of patients seen per shift after AVT implementation.

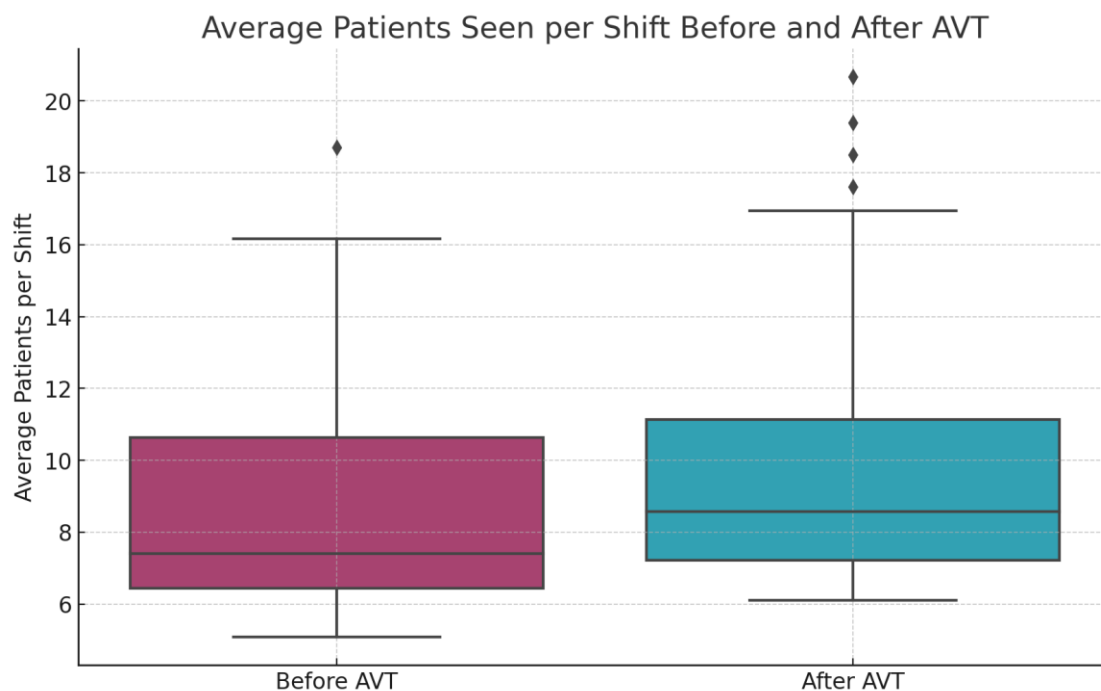


Figure 7.1: Boxplot showing average number of patients seen per clinician shift before and with the AVT

ED - Key Performance Indicator 2

All clinician roles (excluding GPs) experienced a clear and consistent reduction in time to document the first clinical note after the introduction of AVT. The average reduction was 7.02 minutes (range:4.63-11.65 minutes), which was significant ($t = 3.01$; $p<.001$) (Figure 7.2).



Figure 7.2: Boxplot showing the change in documentation times with AVT compared with beforehand

ED - Key Performance Indicator 3

The implementation of AVT led to a measurable and statistically significant improvement in patient flow, shown by a reduced time from clinician assignment to the patient being clinically ready to proceed ($t=3.41$; $p<.001$) (Figure 7.3). The mean % change in improvement was 7.45%.

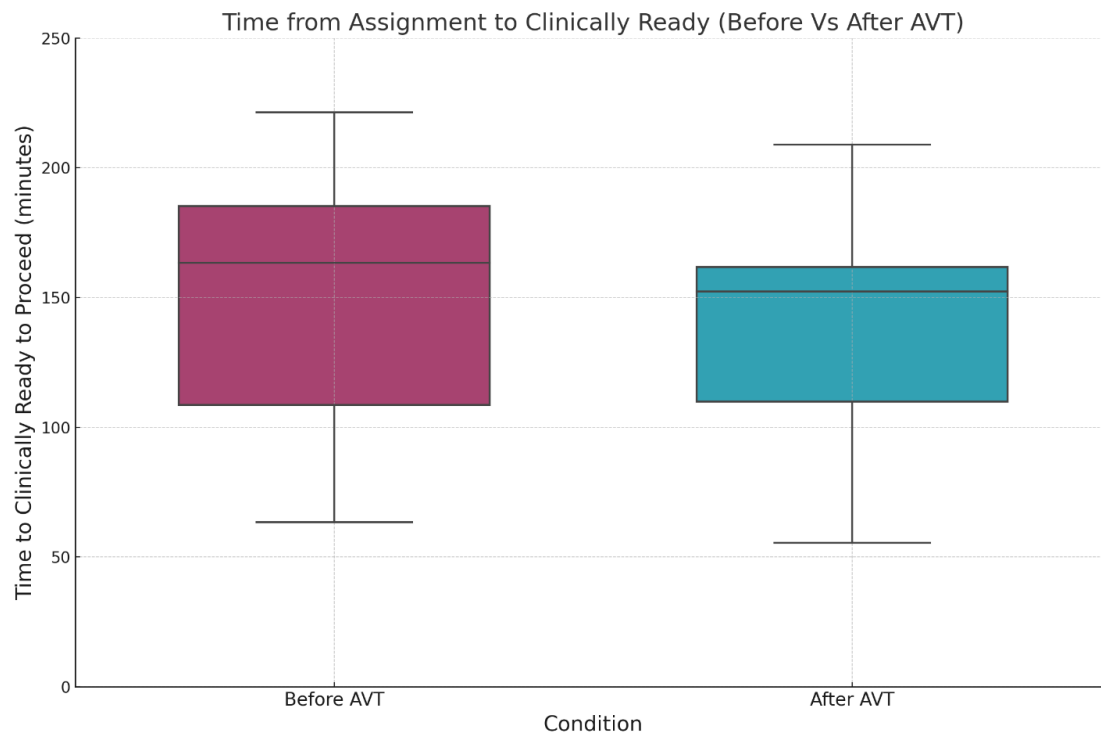


Figure 7.3: Boxplot showing time from assigning to ready to proceed before and after implementation of AVT.

Ambulance Service - Pan-city ambulance service

A second non-core site was a city based ambulance service. The evaluation focused on clinician efficiency in patient assessment processes across two distinct care settings: (1) remote assessments conducted via the Clinical Hub (*Hear and Treat*), and (2) in-person assessments by ambulance-based clinicians (*Face to Face*). Data are presented separately for each setting.

1. Hear and Treat

Implementation

To allow for the usage of AVT within the Clinical Hub (CHUB) working environment an additional piece of hardware was required to merge the audio from the phone to AVT on a workstation computer and was installed on two workstations. Pilot users were instructed to complete their shift as normal with the addition of using AVT to draft their notes for all calls being assessed. There were 11 pilot users, the pattern of their shifts meant that there may not have been a pilot clinician in at times, or at other times there was more than the hardware allowed.

Data Collection and Cohort Description

Quantitative data were collected for the month of April 2025 using existing CHUB performance monitoring tools. Metrics included duration of patient assessments, and the number of patients assessed per hour. AVT users were identified and separated from the CHUB dataset, these clinicians assessed 656 patients (2.7% of total encounters), compared with 23,044 assessments completed by non-AVT users during the same period. No calls were excluded from the Hear and Treat analysis.

Findings

AVT-supported assessments were, on average, completed two minutes faster than those without the tool (15 minutes vs. 17 minutes). In terms of productivity, AVT users assessed 2.3 patients per hour versus 2.0 among non-users, a 15% increase (Figure 7.4).

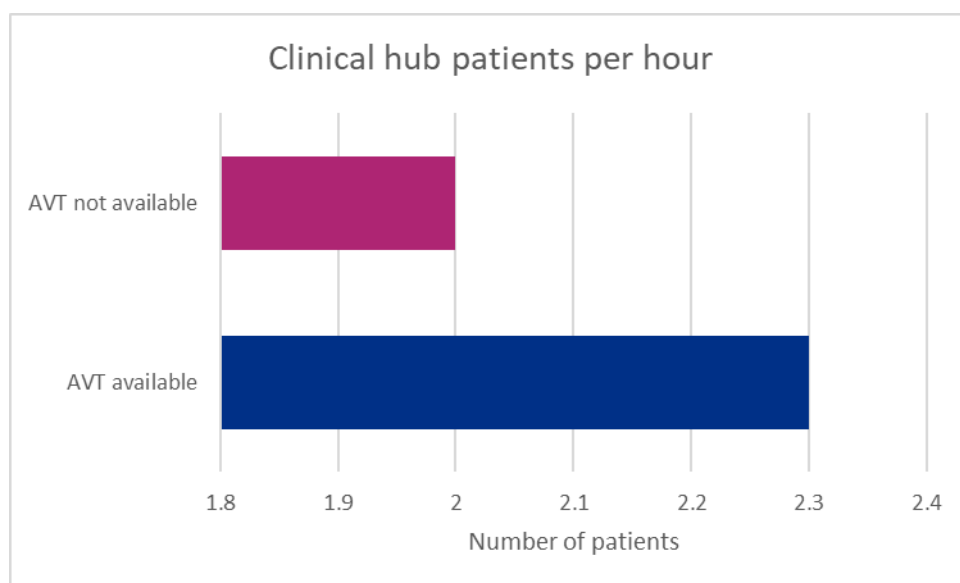


Figure 7.4: Number of patient assessments completed by the Clinical Hub with and without AVT

Documentation Quality

AVT users demonstrated a slight improvement in average documentation quality scores (98.5%) compared with the control group (98.2%), a marginal gain of 0.3%.

2. Face to Face

Data Collection and Cohort Description

A six-week evaluation period was selected and compared to a three-month baseline prior to AVT implementation, which encompassed 816 calls. The cohort included 344 patient contacts involving AVT use, recorded by seven clinicians.

The metrics used were:

- On-scene time (conveyed patient)
- Handover-to-green time (hospital handover to vehicle availability)
- Combined Times (on-scene time + Handover-to-green time)
- See and Treat on-scene time (total time with patient when they are not conveyed to another place for treatment)

On-scene time—the period spent with a patient prior to transport—was reduced by an average of three minutes, representing a 6.8% improvement (44 minutes to 41 minutes).

Handover-to-Green time— (where applicable i.e. for conveyed patients) from hospital handover to vehicle availability—increased marginally from 14 to 15 minutes (value means).

Combined Times—Handover-to-Green and On-scene time merged—showed a reduction of 4.8% or 2.8 minutes in the analysis of the average (58.4 minutes to 55.6 minutes), demonstrating a net productivity gain for the clinicians after the implementation of AVT (Figure 7.5).

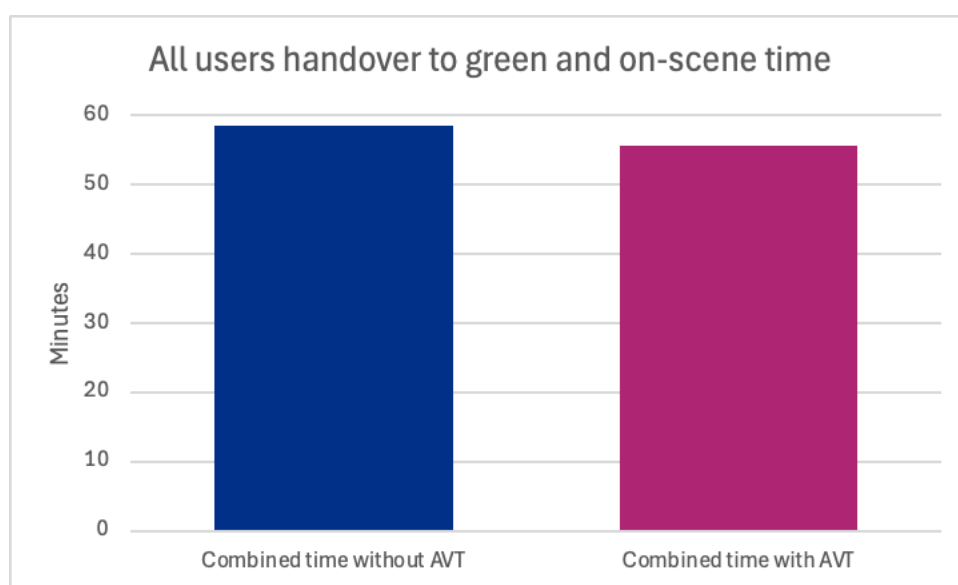


Figure 7.5: All users handover to green and on-scene time combined, with and without AVT

See and Treat Subgroup

A focused analysis of *See and Treat* episodes—where patients are not conveyed to hospital—demonstrated a 3.8% improvement in on-scene time (average reduction of 4 minutes) compared to the pre-AVT evaluation period.

Patients per Shift

An analysis of the number of patients' clinicians were seeing per shift shows that the average for the mean remained similar with a 0.7% reduction in patients seen per shift.

Fast Response Units (FRU)

An additional analysis was conducted of five clinicians participating in the pilot who were assigned to Fast Response Units (FRUs), which are non-conveying assets focused on rapid on-scene assessment and treatment. The evaluation demonstrated

improvements in clinical productivity when using AVT across both mean and median measures of patient throughput.

On average, clinicians using AVT assessed an additional 0.25 patients per shift compared to standard practice (increasing from 4.25 to 4.5 patients per shift), corresponding to a 5.9% improvement. When examining the median number of patients seen per shift, productivity increased from 4 to 5 patients, reflecting a 25% relative improvement. Similarly, the number of patients seen per hour also increased with AVT usage. The mean number of patients seen per hour rose from 0.375 to 0.45—an average increase of 20%. The median improved from 0.35 to 0.45 patients per hour, representing a 28.6% gain in hourly productivity. It is noted that with such a small sample size that there are limitations on the validity and scalability of these data.

7. York Health Economic Consortium (YHEC)



Ambient Voice Technology in Generating Clinical Capacity: ED Summary Results

Introduction

Great Ormond Street Hospital (GOSH) commissioned York Health Economics Consortium (YHEC) to develop a simple calculator to help quantify the potential increase in operational capacity generated using AI technology to record patient consultations.

GOSH examined the impact of this AVT tool in the Emergency Department (ED) department in St George's University Hospital NHS Trust as part of a wider study.

This report focuses on the time saved, how this can increase capacity and the associated opportunity cost benefits created by using the AI tool in the ED setting.

Methods

Data collected as part of this study were used to inform the development of the calculator. As part of the evaluation, 24 ED clinicians used the AVT tool; however the total staff of the ED department is 90. Prior to implementation each clinician saw on average eight patients per shift and spent an average of 12 minutes per patient on documentation such as clinical notes.

After implementation, time spent on documentation tasks reduced by 51.7%, equating to a saving of six minutes per patient. Additionally there was a 13.4% increase in ED capacity per shift per staff member. The analysis assumes that 80% of the time saved can be directly reused by clinicians to see additional patients in the ED department and the average staff member would work a total of 220 days per year accounting for annual leave, training and illness. Therefore one day working per person would equate to one shift.

The calculator also allows users to explore the potential national impact, based on NHS England workforce data (February 2025), which reports 11,055 full-time equivalent ED doctors in England(19).

Cost data were sourced from the Personal Social Services Research Unit (PSSRU) and the 2023/24 NHS National Cost Collection. An average cost of £323.47 (20) per ED attendance and an hourly wage of £93 (21) for emergency medicine doctors were used to estimate the financial value of the released capacity.

Results

Results per Individual

At the individual clinician level, use of the AI tool saved an average of 6 minutes per documentation task, equating to 47 minutes saved per shift. This time saving enabled each ED staff member to see one additional patient per shift. The financial value of time saved on documentation was estimated at £9.33 per task, while the total value of additional capacity created per shift was £270.93. Weekly, this corresponds to £508.68 in documentation time savings and £1,896.53 in added clinical capacity. Annually, this equates to £15,987.19 in documentation time savings and £59,605.10 in additional clinical capacity per individual.

Results per Trust

At a trust level (based on 90 staff), this resulted in a total of 4,219 minutes saved and an additional 20 patient attendances. The corresponding value of time released for documentation tasks was £1,744.06, and the value of additional clinical capacity was estimated at £6,502.37 per shift.

Weekly, this translates to £45,781.50 in documentation time savings and £170,687.32 in added clinical capacity. Annually, the savings total £1,438,847.19 for documentation time savings and £5,364,458.78 for additional capacity across the 90 staff within the trust.

Results per England

At the national level, applying the same assumptions to the full-time equivalent workforce of 11,055 ED doctors resulted in significant cost savings. Total time saved across the workforce per shift was estimated at 518,294 minutes and 9,259 additional ED attendances per day. Weekly, this equates to £5,623,494.44 in documentation time savings and £20,966,093.05 in additional clinical capacity. Annually, the savings amount to £176,738,396.64

for documentation time savings and £658,934,352.89 for added capacity across the national workforce.

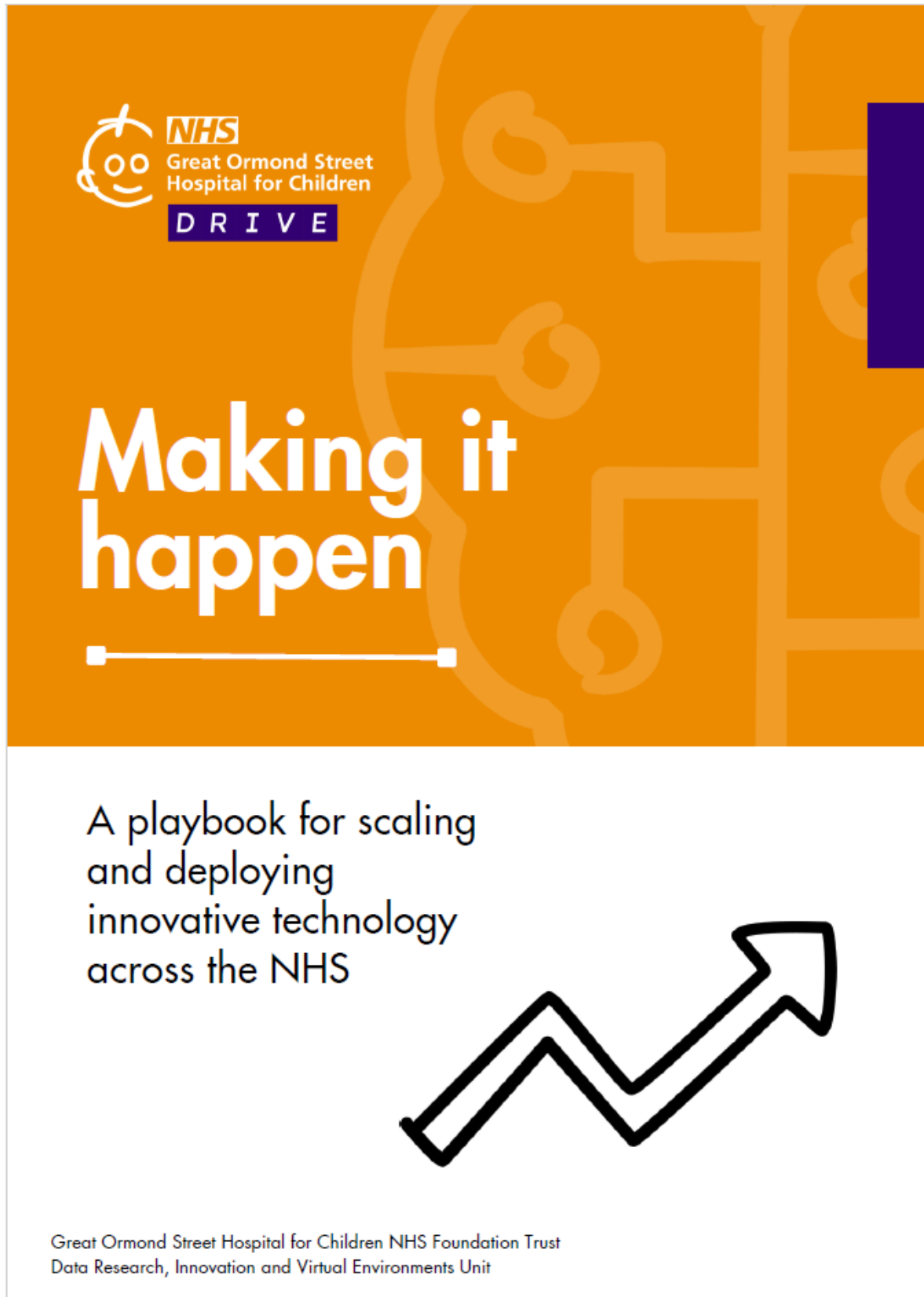
Discussion

The analysis indicates that using ambient voice technology in ED departments could help make services more efficient. The data collected at St. George's Hospital ED showed the AVT tool reduced the time clinicians spent on documentation, creating more time for patient care. If similar results were seen nationally, the additional capacity could support thousands of extra patient attendances each day.

However, there are some important limitations. The results are based on data from just one hospital and a small number of clinicians, so the findings may not apply to all settings. The calculator also relies on a number of assumptions, such as the idea that all time saved can be directly used to see more patients. In reality, how much of this time is actually used for clinical care may vary depending on staff availability, shift patterns, and local pressures.

Further evidence from other sites and over a longer period is needed to understand the full impact and to test whether the benefits are consistent in different environments.

8. Team Learning and Playbook (TBD)



Executive Summary

The NHS is a large consumer of technology and multiple publications and reports have outlined a need for digital innovation. However, technology adoption can be fragmented and variable across the landscape of Integrated Care Systems, Hospital Trusts and regional and national bodies. Lots of ideas are tried; some fail and some succeed, but even when a successful idea comes along, how does it get taken up and spread across NHS organisations – how does it scale?

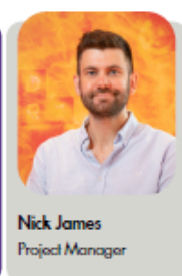
This playbook forms part of a report on a programme sponsored by NHS England London Region and delivered by Great Ormond Street Hospital (GOSH)'s Data Research, Innovation and Virtual Environments (DRIVE) unit, with support from a technology partner, TORTUS AI Ltd. The fifteen-month programme was commissioned to evaluate Ambient Voice Technology (AVT) in a number of clinical settings across London in 2024-5.

Other parts of our programme report provide detailed analysis of the data collected during the evaluation. This playbook is not about those data or the findings of the programme. This playbook is a manual of how to navigate the complex NHS organisational environment to deliver technology projects and programmes at scale. It is based on our experiences during the 15 month programme and assumes you have built a team to deliver your project.

We hope to provide valuable insights and learning that might help other projects – what went well, where we encountered challenges and how we resolved (most) of these.

This playbook is written with NHS colleagues who may be carrying out their own projects, in mind.

We hope you find it useful.



Great Ormond Street Hospital for Children NHS Foundation Trust
Data Research, Innovation and Virtual Environments Unit

MAKING IT HAPPEN

Background



A hub for innovation

GOSH is a leader in healthcare research and innovation, with a dedicated unit for digital innovation, the GOSH DRIVE Unit.

The Innovation Hub at GOSH DRIVE supports development and early-stage testing of data and technology solutions to be scaled in practice at Great Ormond Street Hospital for Children (GOSH) and beyond.

Ambient Voice Technology Pilot

In 2023 GOSH DRIVE signed a collaborative working agreement with a UK based AI Start-up TORTUS.AI (TORTUS), to support the development of their Ambient Voice Technology (AVT) tool. AVT listens to human interactions and uses AI to produce an output. The initial focus was on using the tool to listen to clinical consultations and produce draft clinic notes and (where appropriate) letters.

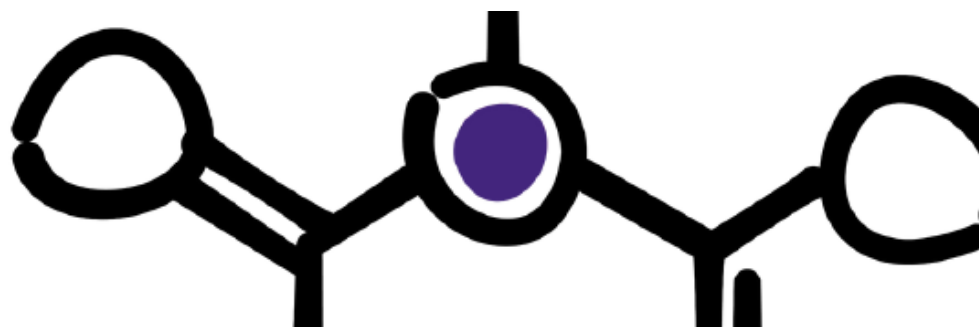
The aim was to evaluate the effectiveness of the tool and its ability to improve the human quality of the consultation, with the clinician multi-tasking less and having the time to focus on the patient, rather than being distracted by typing. The hope was that the technology would also be able to reduce the clinical administrative workload for clinicians, reduce their cognitive load and improve their work-life balance.

Even at early stages, AVT showed a great deal of promise and the team believed that it could deliver significant benefits to the NHS. In the autumn of 2023, senior leaders at GOSH approached the London region of NHS England and suggested that an evaluation be carried out across a number of different healthcare environments to see if AVT provided benefit. An open meeting of Trust and ICB executives was held to demonstrate the technology and gauge interest in joining a pilot.

The project was given the go-ahead in April 2024, with funding from NHS England London region and TORTUS agreeing to provide technology for the evaluation, on the understanding that this was an evaluation of AVT as a type of technology, rather than TORTUS as a provider. It was agreed that the target would be to study 5000 clinical encounters and report after fifteen months.

Great Ormond Street Hospital for Children NHS Foundation Trust
Data Research, Innovation and Virtual Environments Unit

MAKING IT HAPPEN



Hitting the ground running

A senior group began planning how the project would be delivered and commenced recruitment of a team. They also agreed roles and responsibilities. It was agreed that medical staff would be used to observe clinics, as it was felt they would put patients at ease and build rapport with clinicians quickly.

Given budget considerations, only a small team could be recruited and so deployment could only be carried out at one site at a time. These decisions were made before the deployment team was in place to ensure velocity.

As the deployment team started to join, they began liaising with sites who had expressed an interest; in most cases the initial contact was with senior managers. A first site was identified and intensive work carried out to ensure that the project was in a position to go live by the middle of June 2024; with governance complete and engagement, training and observation begun only 10 weeks after project initiation. This was an excellent start.

Meanwhile, engagement had commenced at a number of other sites and a pipeline of where deploy would take place began to firm up.

Deployment - planning meets reality

As roll out at site 1 began, it became clear that more team members would be needed. Clinic schedules for tens of clinicians did not fall into neat compliant patterns that allowed us to maximise our capacity. At times, the team was too stretched to observe all the available clinics, whilst, at other times, there wasn't enough to do.



What was true for observation was also true for clinician training; the flexibility required by clinical colleagues began to put a strain on the resources that the tech partner had available to hold training sessions.



We realised we needed to change course on our deployment team and, as site 1 came to an end, decided to move to using administrative, rather than medical, staff, as we could afford a larger team that way.

We also decided that GOSH NHS staff would train on the technology. This would mean that clinicians would meet and be trained by the same team who would observe clinics. This allowed for rapport and trust to be built more effectively.



By mid-autumn, final sites were being lined up. Despite the experience gained, this involved new learning, as the team began testing AVT in clinical situations that were unfamiliar to most of the team.

At the time of writing (May 2025) data collection has completed across all sites, collecting data on around 15,000 encounters, rather than the 5,000 originally envisaged. From a standing start the team deployed the AVT at eight sites across London, covering GP, Emergency Department, Ambulance, Secondary Care, Community Care, Mental Health and Tertiary Care.

The project findings were reported on time and the project came in under budget.

Great Ormond Street Hospital for Children NHS Foundation Trust
Data Research, Innovation and Virtual Environments Unit

MAKING IT HAPPEN

Engaging the right people

When approaching an NHS organisation with a new project, speak to senior staff to brief them on what you are trying to do. However, ask for a list of contacts who can help you deliver the work. Be clear how soon you need these and that you may not be able to go ahead without them.

When you meet these people, remember, this is their organisation and they know far more about it than you do. Make them your allies in a joint endeavour and be guided by them. If you do this well, they will become your champions. Having internal champions is a must, as they will greatly help in persuading their organisation of the value of your project.

- A Senior Responsible Officer (SRO) to be your escalation point.
- Information Governance (IG) lead.
- ICT contact.
- Communications team contact.
- Relevant clinical leads (Nursing, Medical and Allied Health Professional).
- Operational managers for affected areas and teams.
- If relevant, Patient Experience lead.
- If relevant, a Finance contact.

Begin governance and security work with the IG and ICT (and Finance, if relevant) teams immediately – it can take a lot of time. Your role will be mainly to facilitate conversations between the technology supplier and the local governance teams.

Whilst doing this, you can set up regular project meetings (including an oversight meeting with your SRO) and commence engagement with remaining colleagues. Do work in parallel wherever possible. This is especially true if you are deploying at multiple organisations at once – there is no reason you can't engage five governance teams simultaneously, for example.

What we learnt



Velocity is a must...

Having momentum is crucial. Do your early preparations as soon as possible and get started. Overthinking things often doesn't help. Advance multiple workstreams in parallel; don't queue things up unless you have to.



...but so is flexibility

Getting started quickly is great, but it means your planning may not be perfect. It is very likely that you will need to change something along the way. That something may be major and you may need to change it rapidly (for us it was our staffing model). Make sure your team is ready to accept changes that may need to be made and keep focussed on delivery. If the plan isn't working, it needs to change; agility is key.



Have clear roles and responsibilities

We had very clear roles between GOSH and TORTUS. This really helped avoid confusion. However, we did reallocate them on occasion. An example from our programme was the move to GOSH staff doing the training.



Being an NHS deployment team is such an advantage...

As an NHS team, we understood a lot about the organisations we encountered. Their structure and roles were familiar to us. We were able to build trust quickly and drive engagement, as we were their peers, rather than a company coming out of the blue. This applied to patients, as well as NHS colleagues.



...but it isn't everything

People still have a (very busy) day job. They may not have time to hear about your amazing programme of work. Some people just won't be interested. Engaging isn't easy, but being an NHS colleague will get you a hearing. Show understanding and recognise that you are a guest in their organisation; you will normally have to work around them. The day job comes first!

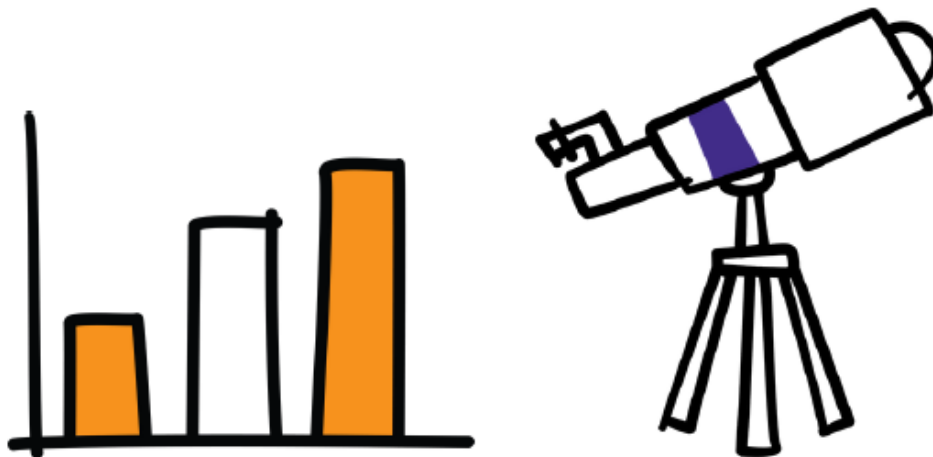
Conclusion

This is an exciting time in healthcare technology, with massive promise to improve the lives of the patients we serve and our hard-working colleagues.

Delivering programmes which bring new innovative technology into large NHS organisations is not easy. There will be things that get in the way of success. A flexible, energetic approach, combined with your NHS knowledge and experience, as well as engaging and working in collaboration with local colleagues, will provide significant advantages in bringing things to a positive outcome – and doing so as quickly as possible.

It is crucial that we do not allow opportunities to improve the service we deliver to fail or wither on the vine. We have a duty to grasp these new technologies and bring them to the front line as speedily and effectively as possible. As change-makers, your role in that is key.


We hope you find this playbook (and our wider report) useful. Thank you for taking the time to read it. The GOSH DRIVE team is always interested in hearing from colleagues in the innovation and health tech fields.



Great Ormond Street Hospital for Children NHS Foundation Trust
Data Research, Innovation and Virtual Environments Unit

MAKING IT HAPPEN



 @GreatOrmondStreetHospitalDRIVE
 www.goshdrive.co.uk

Acknowledgements

GOSH Colleagues

- AVT observations team
- GERALYN Oldham, Quality and Outcomes Manager
- Dr Jo Wray, Senior Research Fellow
- Dr Shankar Sridharan, National Clinical Lead for Artificial Intelligence (NHSE), Chief Clinical Information Officer
- Dr Robert Robinson, Medical Information Officer, Clinical Lead for AI
- Maya Tomlinson, Data Engineer
- Eleanor Sullivan, Communications and Engagement Manager

Research Partner

- TORTUS AI

NHS Pilot Sites

- St George's University Hospitals NHS Foundation Trust
- Crosslands Surgery
- University College London Hospitals NHS Foundation Trust
- North London Mental Health Partnership
- London Ambulance Service NHS Trust
- Kingston and Richmond NHS Foundation Trust

Great Ormond Street Hospital for Children NHS Foundation Trust
Data Research, Innovation and Virtual Environments Unit

MAKING IT HAPPEN

9.NHS T.E.S.T.

At present, there is no consistent national framework within the NHS to assess whether new technologies are clinically effective, safe, and appropriate for use in frontline services. In the absence of clear standards, decisions around adoption can vary widely between organisations. This has, on occasion, led to costly investments that fall short of delivering meaningful change—and, in some instances, may even pose risks to patient safety(22).

In response to the increasing difficulty of assessing digital health technologies, the NHS London Region commissioned the development of the NHS T.E.S.T. framework (Technology Evaluation Safety Test). Designed with real-world use in mind, T.E.S.T. offers a structured yet flexible approach to help ensure that technologies can be safely and effectively scaled across NHS services.

The framework ensures that technologies meet rigorous assurance standards and demonstrate clear evidence of benefit to both patients and healthcare professionals, in line with NHS England guidelines (23).

Technology chosen with consistent evaluation criteria

NHS T.E.S.T. supports the selection of digital technologies that are not only safe and clinically effective, but also practical to scale across a range of care settings. It sits alongside national guidance—such as the NICE health technology evaluation manual—offering a more streamlined, operational tool for frontline use. T.E.S.T. gives healthcare organisations a clear and accessible structure for assessing new technologies, helping to simplify decision-making around implementation and adoption.

The framework adopts a dual-track evaluation approach, assessing both the foundational assurance of a technology platform and the tangible benefits it offers to healthcare delivery:

Platform Assurance (Section A):

Focuses on seven core areas—

- Cybersecurity,
- Data Governance,
- Clinical Safety,
- Bias and Inclusivity,
- Technical Requirements,
- Business Continuity, and
- Emerging Technology. Technologies are expected to meet key standards including NHS DTAC, UK GDPR, and DCB0129 clinical safety requirements.

Benefits Assessment (Section B):

Examines the real-world impact of a solution across 12 domains, ranging from clinical effectiveness and operational efficiency to workforce implications and environmental sustainability. A structured scoring system informs certification levels, helping determine readiness for broader NHS adoption.

The model is flexible and can be tailored to specific technology types seeking national scale or clinical integration. Embedding NHS T.E.S.T. into procurement and digital strategy enables the NHS to drive responsible innovation, make better use of resources, and reinforce confidence among patients and professionals alike.

Supporting scale of AVT

In this case, the T.E.S.T. framework has been applied specifically to Ambient Voice Technology (AVT). However, it can also be extended or adapted to assess other types of digital solutions as new needs arise. The pace at which AI technologies like AVT are being developed and introduced presents challenges around safe implementation and creates uncertainty over how best to select and adopt such tools within the NHS (24). A number of clinicians—and, in some cases, GP practices—have begun using technologies like AVT without confirming whether they deliver proven benefits within the NHS or meet essential standards for information governance and cybersecurity (25). NHS T.E.S.T. is not intended as guidance, but as a practical decision-making tool to determine whether a technology meets the level of assurance and demonstrated benefit required for wider adoption. Technologies being considered for national scaling must meet higher standards than those still in early development or limited local use.

The framework places particular emphasis on clinical effectiveness, cost-effectiveness, and workforce impact. At its core, NHS T.E.S.T. is designed to ensure that new technologies lead to genuine improvements in care. Without robust, real-world evidence of benefit, there is a risk that technologies will be introduced without delivering meaningful value to patients. Solutions seeking national uptake should be supported by rigorous evaluation—such as randomised controlled trials, large-scale studies, or well-powered NHS pilots. Only those with compelling evidence of impact qualify for Gold Certification, indicating their readiness for broader NHS deployment. This level of scrutiny helps to reduce the risk of wasted investment and avoidable patient safety issues.

Ensuring safety, effectiveness, and sustainability is essential before any new system is deployed at scale. While there are often concerns that regulation can slow innovation, NHS T.E.S.T. is designed to do the opposite: by setting out clear, practical benchmarks for assurance, it provides a structured route to NHS adoption. For suppliers, this removes uncertainty around approval and procurement and supports a more efficient development pipeline. Whether for start-ups or established

vendors, the framework encourages the design of technologies that are truly fit for NHS use.

Importantly, NHS T.E.S.T. is not intended to limit local innovation or clinician choice. Rather, it establishes a consistent, evidence-based approach to evaluating new technologies—helping to ensure that the tools we invest in genuinely work, deliver benefit, and are ready to support care at scale.

As the digital landscape continues to evolve, NHS T.E.S.T. will be reviewed and refined to reflect emerging clinical priorities and the pace of innovation. Longitudinal studies will play a key role in assessing the long-term impact of T.E.S.T.-approved technologies on patient care and system efficiency

[Further information on T.E.S.T.](#)

<https://healthinnovationnetwork.com/resources/nhs-test-an-intelligent-framework-to-choose-new-technologies/>

Access the NHS TEST Framework here to find out more about how you can use it.

<https://healthinnovationnetwork.com/wp-content/uploads/2025/06/NHS-T.E.S.T-Part-A-and-Part-B-summary-V11.17625.SS.pdf>

10. Discussion

Summary of results

Results from all sources of data collection indicate strong support for the use of AVT in clinical settings. There was a 16.5% increase in direct care across core sites and a 8.2% reduction in the total time of appointments. Use of AVT in the emergency department resulted in significant increases in the numbers of patients seen in an ED shift, a significant reduction in the time taken to complete ED documentation and significant improvements in patient flow. For the ambulance service, AVT improved clinical efficiency across both remote and in-person care contexts: within the Clinical Hub environment, clinicians completed telephone assessments more rapidly and managed higher patient volumes without compromising documentation quality; in an ambulance operational environment combined on-scene and handover-to-green times were reduced. Clinician feedback indicated significantly more positive scores with AVT on questions related to time and attention given to patients, how distracting, stressful and disruptive computer tasks were, satisfaction with care given and overall experience and a higher level of satisfaction with both effort and time taken to check and edit clinic notes and letters. There was a significant reduction in total NASA cognitive load score in the AVT condition compared with baseline.

The qualitative data (free text comments and interviews) indicated that enthusiasm for the AVT spans all clinical contexts and it is seen as a potentially transformative tool in clinical practice. Benefits in terms of workload, cognitive load and emotional wellbeing were widely reported, corroborating the reduction in total NASA score. Of particular note was the identified value of AVT to neurodivergent clinicians. Benefits to patients were reported by clinicians in terms of improved consultation experience and indirect improvements in clinical outcomes.

Some concerns and challenges were raised in relation to the templates in particular and the loss of some of the more nuanced elements of a consultation, including the clinician voice and clinical details for patients with multiple diagnoses. Whilst some clinicians mentioned improved productivity in terms of numbers of patients seen, concern was also raised about expectations to see more patients as a result of improved efficiency, particularly in services for patients with complex needs.

Whilst there were no significant differences in patient or parent/carer experiences with and without AVT, with experience being generally very positive, comments indicated improved interactions with clinicians when AVT was being used and a high level of acceptance of it. Net promoter scores were favourable.

Clinicians were clear that successful implementation will be contingent on setting-specific adaptations, robust template training and design, system integration, and alignment with clinician documentation standards. Customisation, accessibility, and flexibility remain essential for scalability across the diverse landscapes of healthcare practice.

Strengths and Limitations

Our clinical evaluation study is the first in the NHS to test AVT in a range of diverse clinical settings and evaluate it using both objective and subjective data collected from multiple sources. In addition, we have developed a playbook for the implementation of AVT in terms of scaling and deploying it across the NHS, worked with YHEC to develop a simple calculator to help quantify the potential increase in operational capacity generated using AVT and developed the NHS T.E.S.T. framework, a structured yet flexible approach to help ensure that technologies can be safely and effectively scaled across NHS services in a real-world setting.

Of note, our evaluation happened over a 12 month period during which the training, templates and hardware workarounds were constantly evolving. A strength was the collaboration with the technology company and their responsiveness to challenges and concerns as they emerged, resulting in a product which improved over time (in real-time). This was also reflected in clinicians' experiences and neatly summarised by the low net-promoter score for the AVT from one of the first organisations to be recruited to the trial compared with scores from sites which participated at later stages.

There are some limitations which need to be considered in the context of the findings:

- Participating clinicians self-selected
- Not all professions/specialties represented
- Lack of paired survey data for patients/parents/carers
- Use of unvalidated surveys for clinicians, patients and parents/carers
- TimeCat data could not be collected at non-core sites – but operational data were collected, providing indicators of efficiency/productivity
- For some sites there was a need for some technology workarounds (particularly in relation to hardware), with resulting delays and potentially less data capture
- AVT was not integrated with all EPRs, which impacted clinician experience even though this was not a reflection of the AVT *per se*

Although this study incorporated a wide range of clinical settings, it is still important to be cautious when extrapolating results to other sites, particularly over extended time periods. All clinical settings involve diverse and complex interactions and processes, and the full impact of any change may only become apparent in the longer term.

Next steps

Data-driven healthcare delivery has long been an aspiration across the NHS, but, too often, ambition has outpaced implementation. The Ambient Voice Technology (AVT) Phase 4 evaluation has demonstrated that this need not be the case. This was the first scientifically rigorous, multi-site NHS-led evaluation of AVT, which not only confirmed the technology's clinical utility, but provided a blueprint for how AI can be adopted safely, at scale, and with real-world impact.

The findings are already shaping national direction: they have informed NHS England's official guidance on AI-enabled scribing, contributed to the NHS Spending Review, and are embedded in the Government's forthcoming 10-year plan for health innovation and productivity. Crucially, the work has highlighted that success depends not only on the AI itself, but on aligning people, processes and platforms. Through coordinated governance, frontline training, workflow redesign and ongoing feedback loops, AVT improved documentation quality, patient-clinician interaction, and clinician wellbeing, while unlocking productivity and efficiency gains across an Emergency Department and other settings.

This programme has shown that the NHS can evaluate emerging technologies with scientific rigour, pace, and strategic impact. But to fully realise AVT's full benefit, it must be deployed as more than a standalone tool. It should be treated as a platform capability. A 'platform play' approach consolidates value, reduces duplication, and enables consistency, allowing AVT to evolve from a point solution into an engine for clinical and operational reform.

With NHS T.E.S.T. now established as a national framework for selecting assured tools of proven benefit, the NHS has both the method and the momentum to act. This is not about narrowing choice of vendor. Rather, it is about ensuring that technologies selected for widespread use are safe, evidence-based, and deliver proven value. Approved solutions must be able to share structured data and operational insights, enabling the NHS to improve national visibility, enhance data quality, and unlock system-wide intelligence.

The methodology used in this evaluation provides a replicable model for assessing other high-potential technologies the NHS may wish to scale. NHS T.E.S.T., developed to support this programme, is a practical and adaptable evaluation framework that can be applied across a wide range of innovations, helping to guide future decisions, based on assurance, benefit, and fitness for purpose.

The priority now is to strengthen strategic deployment. By investing in coordinated, assured, and evidence-led implementation, the NHS can scale AVT with confidence improving care for patients, reducing the burden on staff, and helping the system operate more effectively. The groundwork has been laid. What comes next is delivery.

The priority now is to strengthen strategic deployment. By investing in coordinated, assured, and evidence-led implementation, the NHS can scale AVT with confidence, improving care for patients, reducing burden on staff, and helping the system operate more effectively. The groundwork has been laid. What comes next is delivery.

11. Appendices

Appendix A – example information leaflet text

Taking part in the Teddington Memorial TORTUS Ambient AI trial

Introduction

In your upcoming appointment, your clinician may be using a new computer-based tool to help them write their clinic note and letter (if relevant) about your visit. The tool is called TORTUS - it listens to the conversations during the appointment and generates a note summary and clinic letter (if relevant) at the end. This could allow the clinician to spend less time taking notes or typing on the computer and more time focusing on you.

Hounslow and Richmond Community Healthcare NHS Trust (HRCH) at Teddington Memorial Hospital is working in partnership with Great Ormond Street Hospital (GOSH) to evaluate TORTUS.

Why is TORTUS being used in my appointment?

HRCH at Teddington Memorial Hospital is trialing TORTUS in some clinics. We want to see whether TORTUS has any effect on:

- the length of the appointment
- the quality of the interaction between the clinician and you.
- the experiences of you and the clinician.

How does TORTUS listen to what I am saying?

TORTUS uses 'ambient' technology. This means it can turn words spoken by you and the clinician into written words (text).

How does TORTUS turn my words into the clinician's note and letter?

TORTUS uses artificial intelligence (AI), which has been tested by clinicians. TORTUS has been programmed to listen to all of the words spoken during the appointment. At the end of the appointment, TORTUS summarises the important clinical content in a clinic note and letter (if relevant). The clinician will then check both for accuracy and completeness before making any changes and approving them.

Is TORTUS safe to use in my appointment?

We know that some people have concerns about the use of AI technology. Before this trial could take place, HRCH completed detailed and thorough safety checks to ensure that the technology is secure.

Does TORTUS keep my words?

No - TORTUS will not keep your words. All data are kept in secure environments and are deleted after your notes and letter (if relevant) are generated at the end of the consultation. TORTUS makes a summary note and letter (if relevant) from the words spoken during the appointment. These will be copied into your record by the clinician. Only the clinician-

approved clinic note and clinic letter (if relevant) will exist in the electronic patient record after the clinic.

Will my appointment be different because TORTUS is being used?

Your appointment format will not change. The only differences are that you will be asked if you are happy to consent to TORTUS being used before meeting the clinician.

With your written consent, there will be a member of GOSH staff in the room, observing the appointment. They will be there to record how long the appointment takes and how long the clinician spends talking to you.

What if I change my mind?

The observer will immediately leave the appointment if you, or your clinician, ask them to.

Do I need to do anything extra for my clinic appointment?

No, you do not need to do anything extra before your appointment.

Will I be asked to do anything after my clinic appointment?

We will invite you to complete a short survey about your experience of your appointment. The survey should take no more than five minutes.

Will TORTUS be used in my future appointments?

Once the TORTUS trial has finished, we will review all of the findings. If the findings suggest that patients and clinicians have a better experience in clinic with TORTUS compared to clinics without TORTUS, HRCH will explore whether we should introduce TORTUS across Teddington Memorial Hospital outpatient clinics in the future.



What if I have more questions?

If you have questions about the use of TORTUS in your appointment, please ask a member of the GOSH project team who will be present and available during your visit.

Participation

Your participation is voluntary. You are free to decline to participate at any time, without giving any reason, without your medical care or legal rights being affected.

Taking part in the Teddington Memorial Hospital TORTUS trial

Introduction

In the upcoming appointment, your clinician may be using a new computer-based tool to help them write a clinic note and letter (if relevant) about your child's visit. The tool is called TORTUS - it listens to the conversations during the appointment and generates a clinical summary and letter (if relevant) at the end. This could allow the clinician to spend less time taking notes or typing on the computer and more time focusing on you and your child.

Hounslow and Richmond Community Healthcare NHS Trust (HRCH) at Teddington Memorial Hospital is working in partnership with Great Ormond Street Hospital (GOSH) to evaluate TORTUS.

Why is TORTUS being used in my child's appointment?

HRCH at Teddington Memorial Hospital is trialing TORTUS in some clinics. We want to see whether TORTUS has any effect on:

- the length of the appointment
- the quality of the interaction between the clinician and you and your child • the experiences of you, your child, and the clinician.

How does TORTUS listen to what I am saying?

TORTUS uses 'ambient' technology. This means it can turn words spoken by you, your child and the clinician, into written words (text).

How does TORTUS turn my words into the clinician's note and letter?

TORTUS uses artificial intelligence (AI), which has been tested by clinicians. TORTUS has been programmed to listen to all of the words spoken during the appointment. At the end of the appointment, TORTUS summarises the important clinical content in a clinic note and letter (if relevant). The clinician will then check both for accuracy and completeness before making any changes and approving them.

Is TORTUS safe to use in my appointment?

We know that some people have concerns about the use of AI technology. Before this trial could take place, HRCH completed detailed and thorough safety checks to ensure that the technology is secure.

Does TORTUS keep my words?

No - TORTUS will not keep your words. All data are kept in secure environments and are deleted after your notes and letter (if relevant) are generated at the end of the consultation. TORTUS makes a summary note and letter (if relevant) from the words spoken during the

appointment. These will be copied into your child's record by the clinician. Only the clinician-approved clinic note and clinic letter (if relevant) will exist in the electronic patient record after the clinic.

Will my appointment be different because TORTUS is being used?

Your appointment format will not change. The only differences are that you will be asked if you are happy to consent to TORTUS being used before meeting the clinician.

With your written consent, there will be a member of GOSH staff in the room, observing the appointment. They will be there to record how long the appointment takes and how long the clinician spends talking to you and your child.

What if I change my mind?

The observer will immediately leave the appointment if you, your child, or your clinician ask them to.

Do I need to do anything extra for my clinic appointment?

No, you do not need to do anything extra before your appointment. If your child is old enough, it may be helpful to talk to them about the clinician using TORTUS at their next appointment.

Will I be asked to do anything after my clinic appointment?

We will invite you and your child to complete a short survey about your experience of your appointment. The survey should take no more than five minutes.

Will TORTUS be used in my future appointments?

Once the TORTUS trial has finished, we will review all of the findings. If the findings suggest that patients, families, and clinicians have a better experience in clinic with TORTUS compared to clinics without TORTUS, we will explore whether we should introduce TORTUS across Teddington Memorial Hospital outpatient clinics in the future.



What if I have more questions?

If you have questions about the use of TORTUS in your appointment, please ask GOSH project team staff who will be present and available during your visit.

Participation

Your and your child's participation are voluntary. You are free to decline to participate at any time, without giving any reason, without your child's medical care or your child's legal rights being affected.

Appendix B – example consent forms

Adult Patient Consent Form

Participation in a Service Evaluation of TORTUS in Clinical Practice

MRN	Clinic Date	Clinician Name
<p>1. I confirm that I have read and understand the information provided about the TORTUS Ambient AI project.</p> <p>2. I have had the opportunity to consider the information, ask questions and I have had these answered satisfactorily.</p> <p>3. I understand that:</p> <ul style="list-style-type: none"> • My participation in the service evaluation of TORTUS is voluntary and that I am free to withdraw from use of the TORTUS tool at any time during the consultation, without giving any reason and without my medical care or my legal rights being affected. • I understand that the evaluation is a joint project between Great Ormond Street Hospital (GOSH) and University College London Hospitals (UCLH). • The clinic appointment in which I am participating will be observed by GOSH staff. I agree to my clinic appointment being observed. • The observer/s will leave the room if asked to by a clinician. • I can request the observer/s to leave the clinic room at any time during the clinic appointment. • TORTUS AI will listen to the conversation between me, and the clinician during the appointment today. • At the end of today's appointment TORTUS AI will summarise the conversation and will capture the important clinical content in a clinic note and a clinic letter. • The clinician will check the clinic note and/or clinic letter for accuracy and completeness before approving it and storing it in my electronic health record held at UCLH. • No personal data will be held by any systems outside of the UCLH electronic patient record once the note and letter are generated. <p>4. I agree to take part in the service evaluation of TORTUS AI as outlined above.</p>		

Patient Forename

and

Patient Surname

Date

Signature of Patient

Name of GOSH Staff taking consent

Paediatric Patient Consent Form

Participation in a project studying clinicians and their engagement with TORTUS Ambient AI during clinical appointments

MRN	Clinic Date	Clinician Name
<p>1. I confirm that I have read and understand the information provided about the TORTUS project.</p> <p>2. I have met with a member of the Great Ormond Street Staff (GOSH) project team:</p> <p style="margin-left: 40px;">a. My child and I have had the opportunity to ask questions.</p> <p style="margin-left: 40px;">b. My child and I have had our questions answered satisfactorily.</p> <p>3. I understand that:</p> <ul style="list-style-type: none"> • My child's and my participation in the evaluation of TORTUS is voluntary and that I / we are free to withdraw at any time, without giving any reason and without my child's medical care or my child's legal rights being affected. • I understand that the evaluation is a joint project between Great Ormond Street Hospital (GOSH) and University College London Hospitals (UCLH). • The clinic appointment in which my child and I are participating in will be observed by GOSH staff. I agree to my child's clinic appointment being observed. • The observer/s will be timing the tasks that the clinician is carrying out in the appointment. • I can ask the observer/s to leave the clinic room at any time during the clinic appointment. • The observer/s will leave the room if asked to by the clinician. • TORTUS AI will listen to the conversation between me/my child, and the clinician during the appointment today. • At the end of today's appointment TORTUS AI will summarise the conversation and will capture the important clinical content in an appointment note and/or letter. • My clinician will check the appointment note and/or letter for accuracy and completeness before approving it and storing it in my electronic health record held at UCLH. • None of my child's personal data will be held by any organisation or electronic system outside of the UCLH once the appointment note and/or letter are generated. • At the end of my appointment, I will be invited to complete a short survey about my clinic appointment experience. <p>4. I agree to take part in the service evaluation of TORTUS as outlined above.</p>		

Patient Forename and Patient Surname Date

Signature of Parent / Carer

Name of GOSH Staff taking consent

Appendix C – example surveys

Example BASELINE surveys

CLINICIAN



Baseline Clinician Survey

PARENT-CARER



Baseline Parent-Carer Survey

PATIENT



Baseline Adult Patient Survey

Example AVT Surveys

CLINICIAN



ED Clinician – AVT Survey

PARENT-CARER



AVT Parent-Carer Survey

PATIENT



AVT Adult Patient Survey

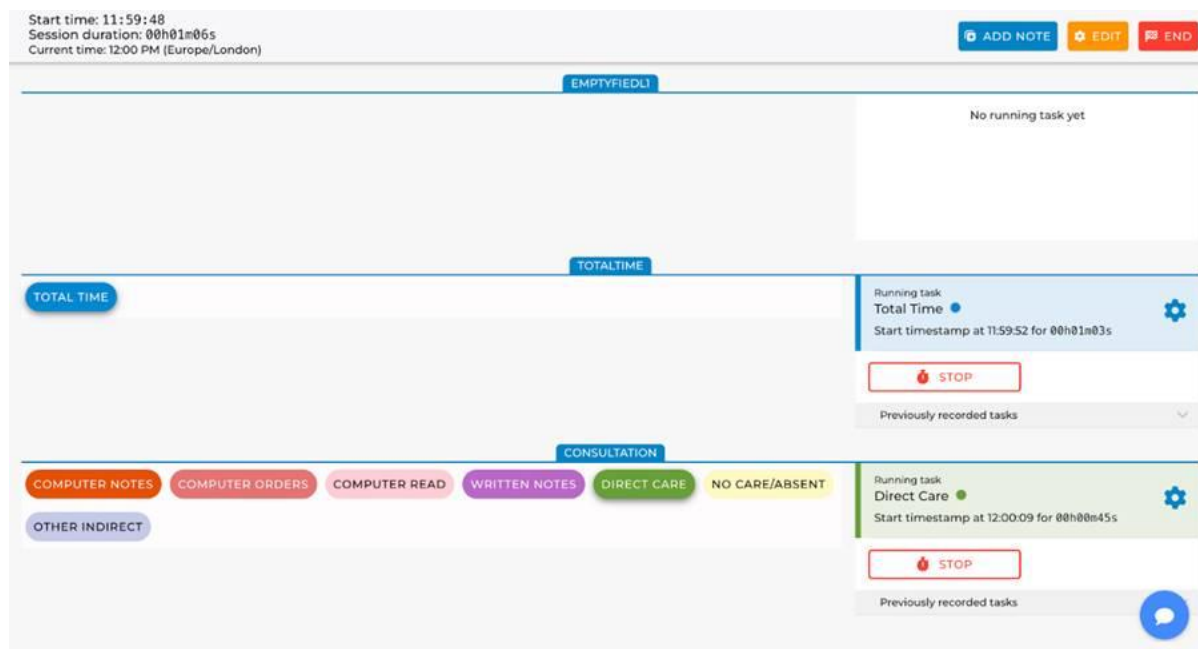
Appendix D – interview example topic guide

Phase 4 – AVT Semi structured Interviews - Topic Guide

<p>What was your overall experience of using Tortus AVT?</p> <p>Intro</p> <ul style="list-style-type: none"> • The purpose of this session is to understand your experience of using ambient AI. • What is good? What didn't work, what do you like or not like? • Be as honest as you want, we will not share the interview transcripts, and any reports will be anonymised.
<p>Can you describe your usual practice?</p> <p>What did you hope that AVT would do for you, what did you expect from it?</p> <p>What was your experience of using AVT?</p> <p>Did AVT meet your expectations?</p>
<p>Thinking first about this technology - How confident did you feel in the AVT technology? Did you feel you could trust the technology?</p> <p>If you trusted it – was your trust well placed?</p>
<p>How do you think that using Tortus influenced or impacted your interactions with the patient OR patient and family?</p>
<p>What was your overall impression of the final outputs – your NOTE?</p> <p>Did you read or use the transcript (discriminate between transcript and notes/letters) (accuracy, hallucinations, etc)</p>
<p>Training and Templates – was training adequate, did the templates work, did you change them?</p>
<p>Do you think AVT would have any impact on patient / clinical outcomes?</p>
<p>If there is time at the end of the interview, ask participants:</p> <p>Is there anything we have not covered or anything else you would like to add?</p>

Appendix E: Additional quantitative (TimeCat) data analysis

Screenshot of TimeCaT tool



Exploratory Analysis

Fixed Effects

Plots were created to explore the distribution of contributing variables across observation types to investigate the potential for confounding variables. As evidenced in the plots below (Figures E.1, E.2 and E.3), all fixed variable distributions were similar across arms. This balance helps ensure that observed outcome differences can be more confidently attributed to the use of AVT.

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings
Across the NHS

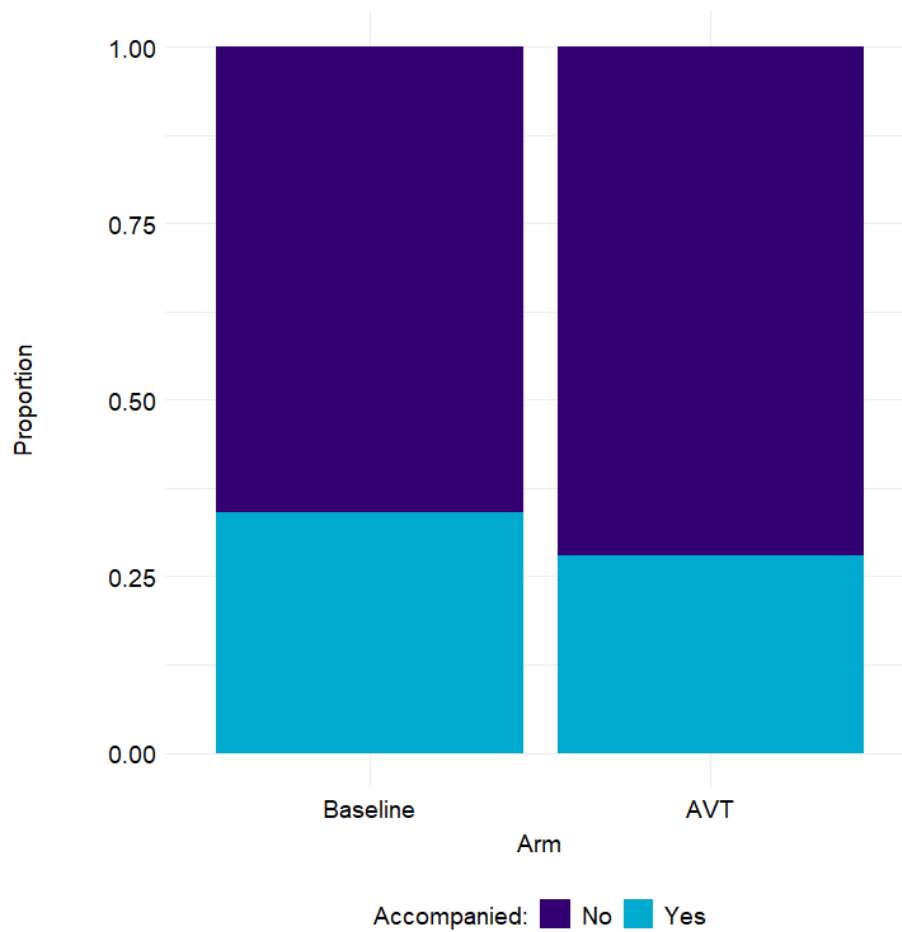


Figure E.1.: Distribution of the proportion of accompanied status across Baseline and AVT arms

The use of Ambient Voice Technology with Generative Artificial Intelligence in Multiple Clinical Settings
Across the NHS

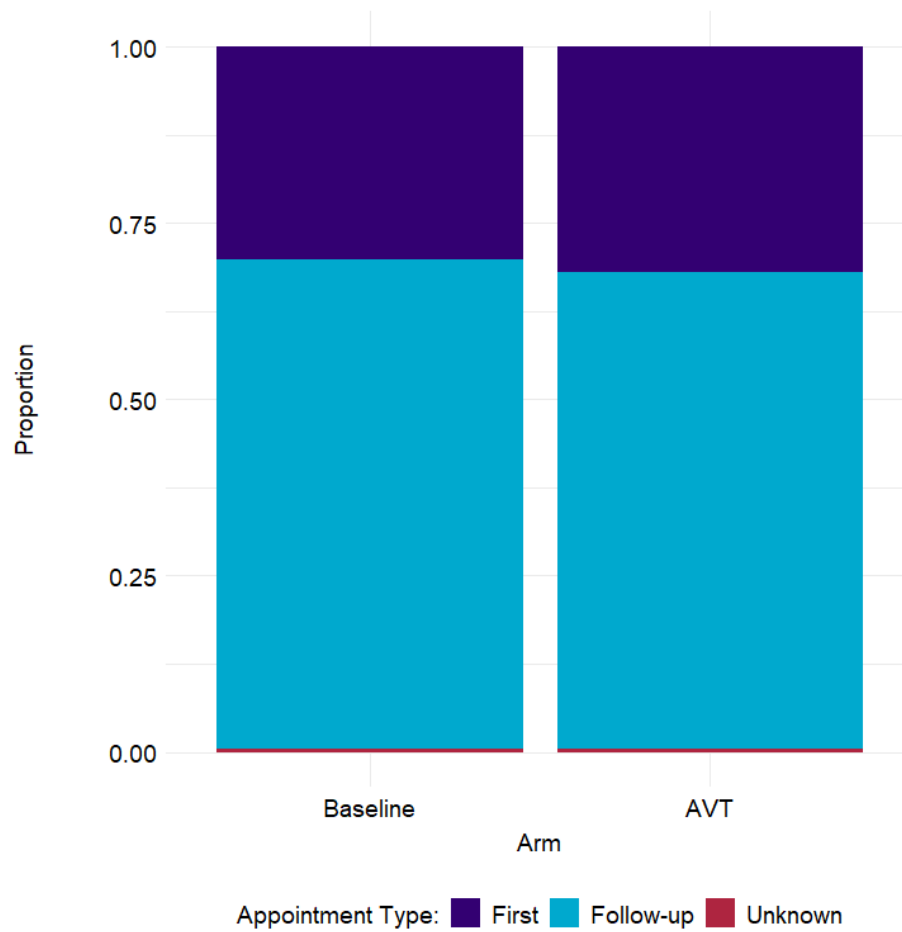


Figure E.2: Distribution of the proportion of appointment types across Baseline and AVT arms

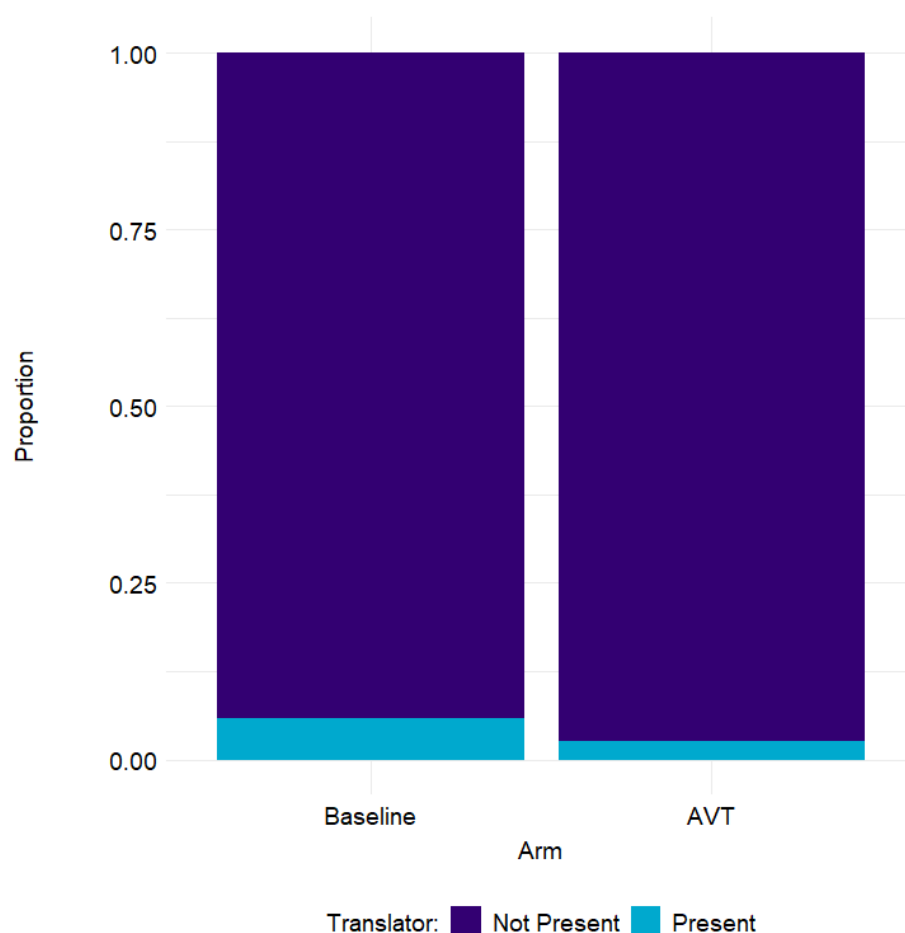


Figure E.3: Distribution of the proportion of translator status across Baseline and AVT arm

Random Effects

Plots were created to explore the distribution of direct care and total time at each arm across sites to investigate the variability between these variables. As evidenced in the plots below (Figures E.4, E.5, E.6 and E.7), it is clear there is variation across sites and clinicians. This is important to consider within analysis because the effect of AVT will differ across these levels.

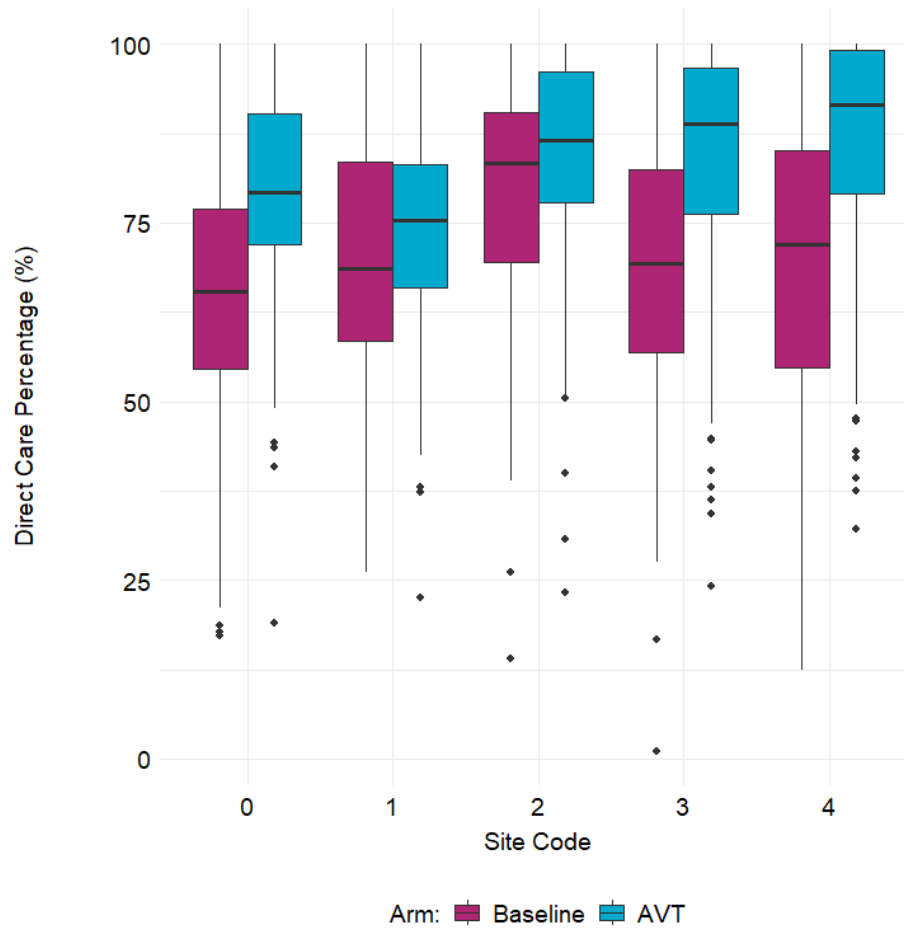


Figure E.4: Variation in percentage of session time spent on direct care between baseline and AVT arm across sites

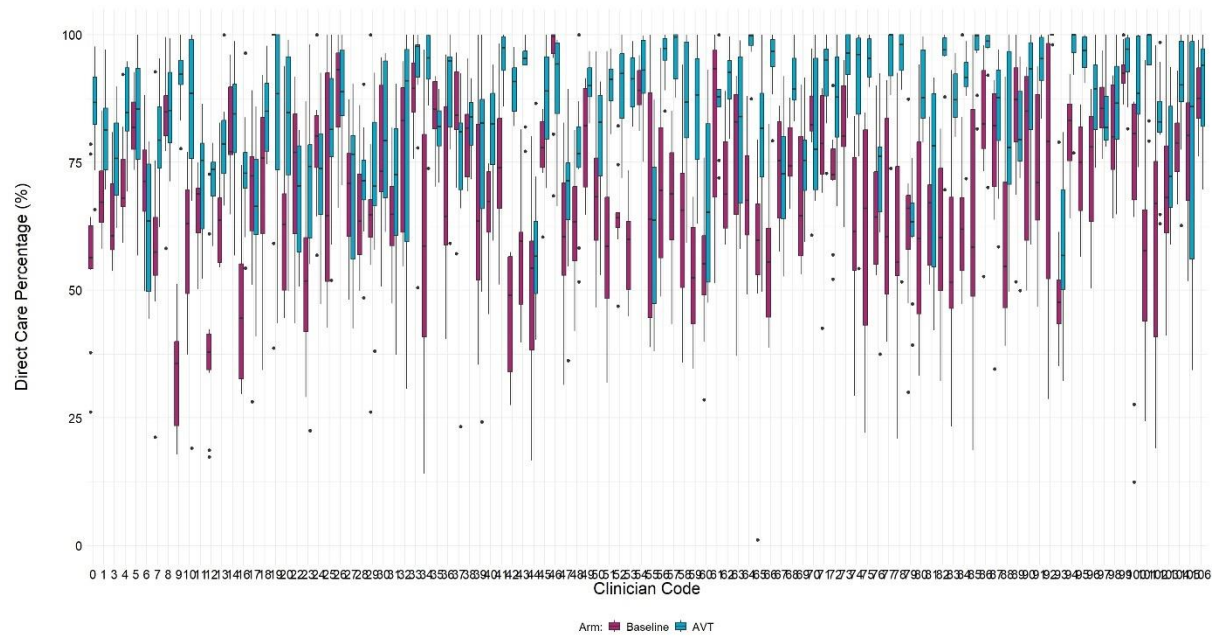


Figure E.5: Variation in percentage of session time spent on direct care between baseline and AVT arm across clinicians

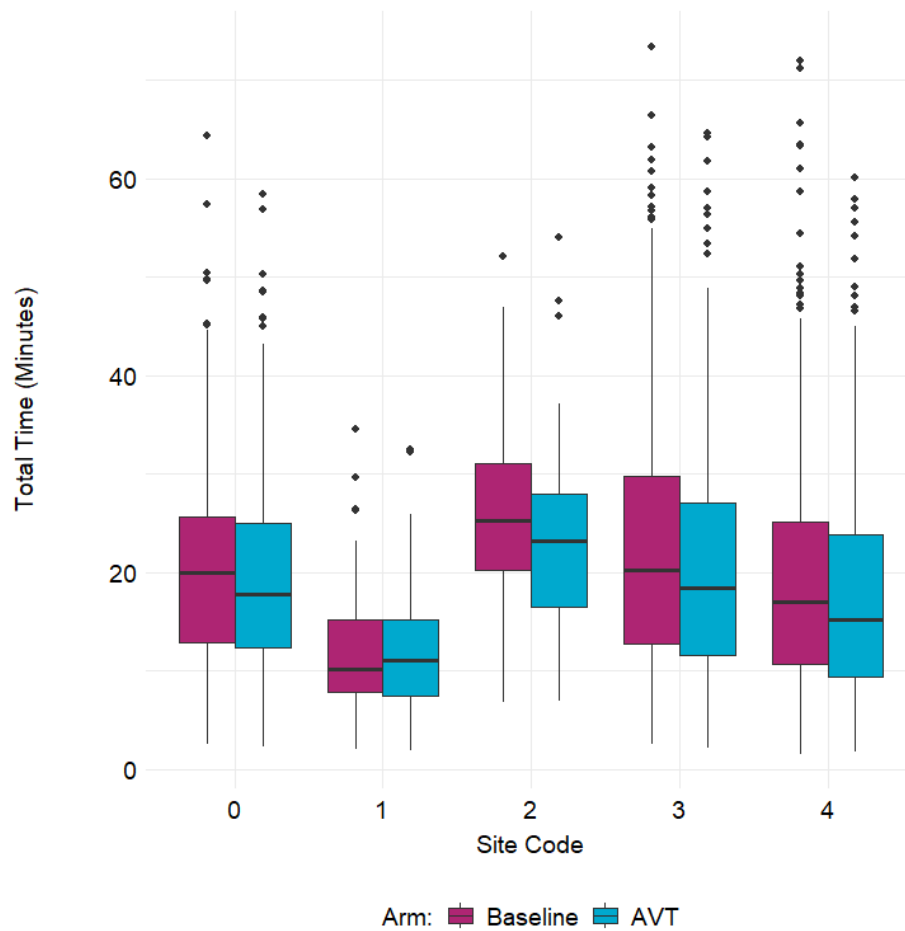


Figure E.6: Variation in total session time between baseline and AVT arm across sites

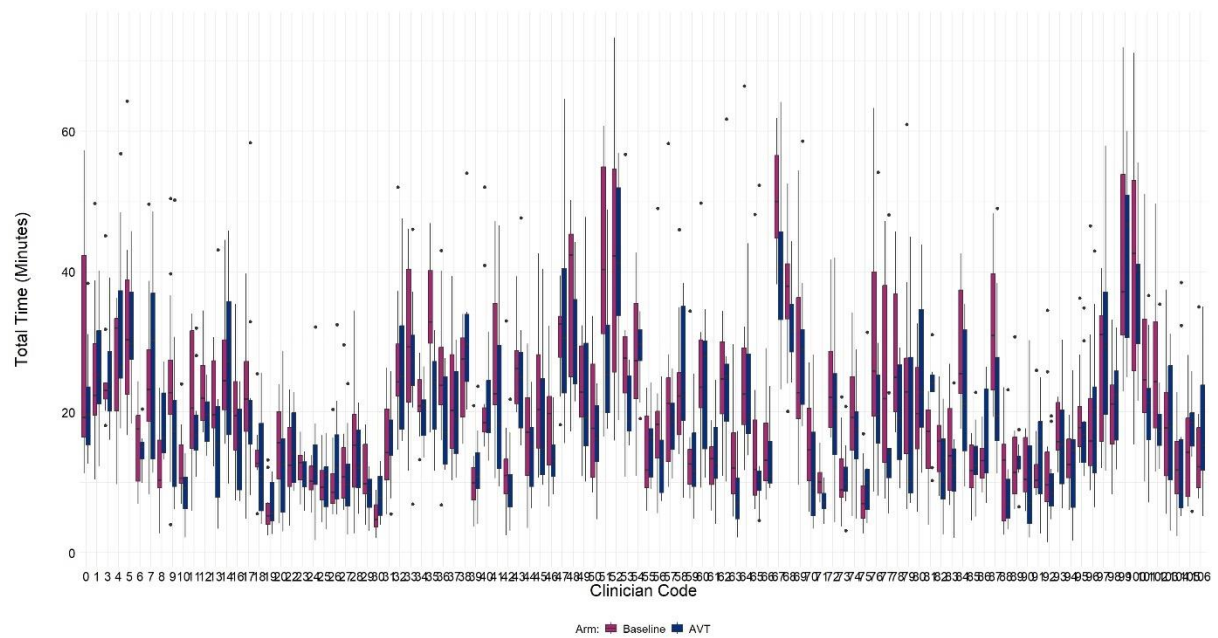


Figure E.7: Variation in total session time between baseline and AVT arm across clinicians

Dependent Variables

Plots of the distribution of direct care and total time variables were created to help determine the appropriate statistical method and verify assumptions for analysis.

Direct Care

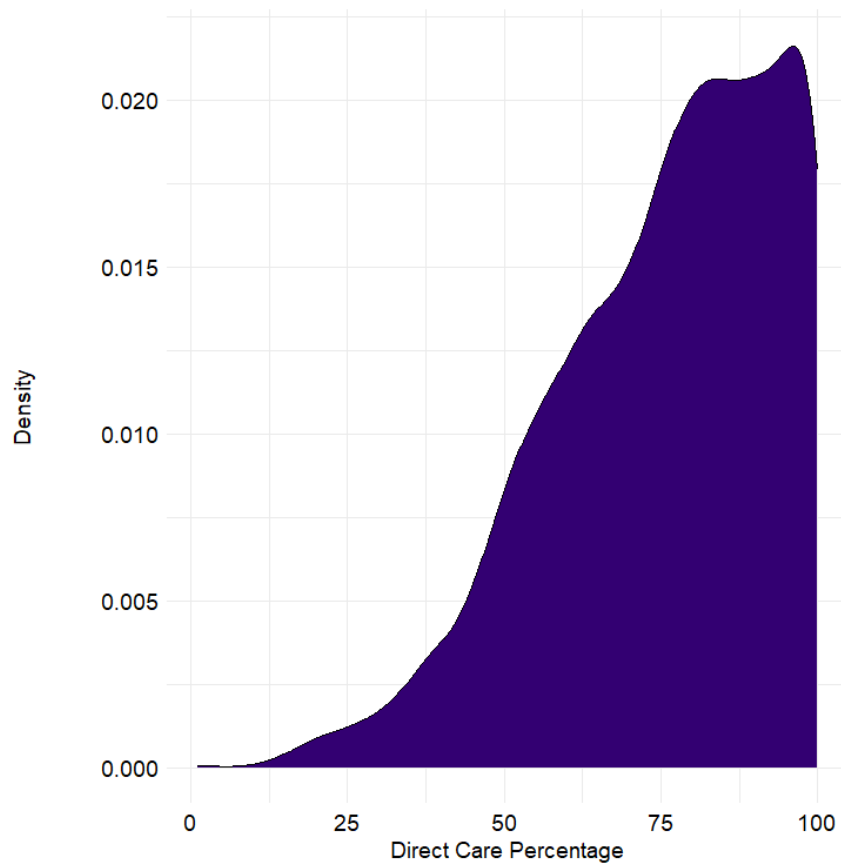


Figure E.8: Density plot showing distribution of direct care percentage

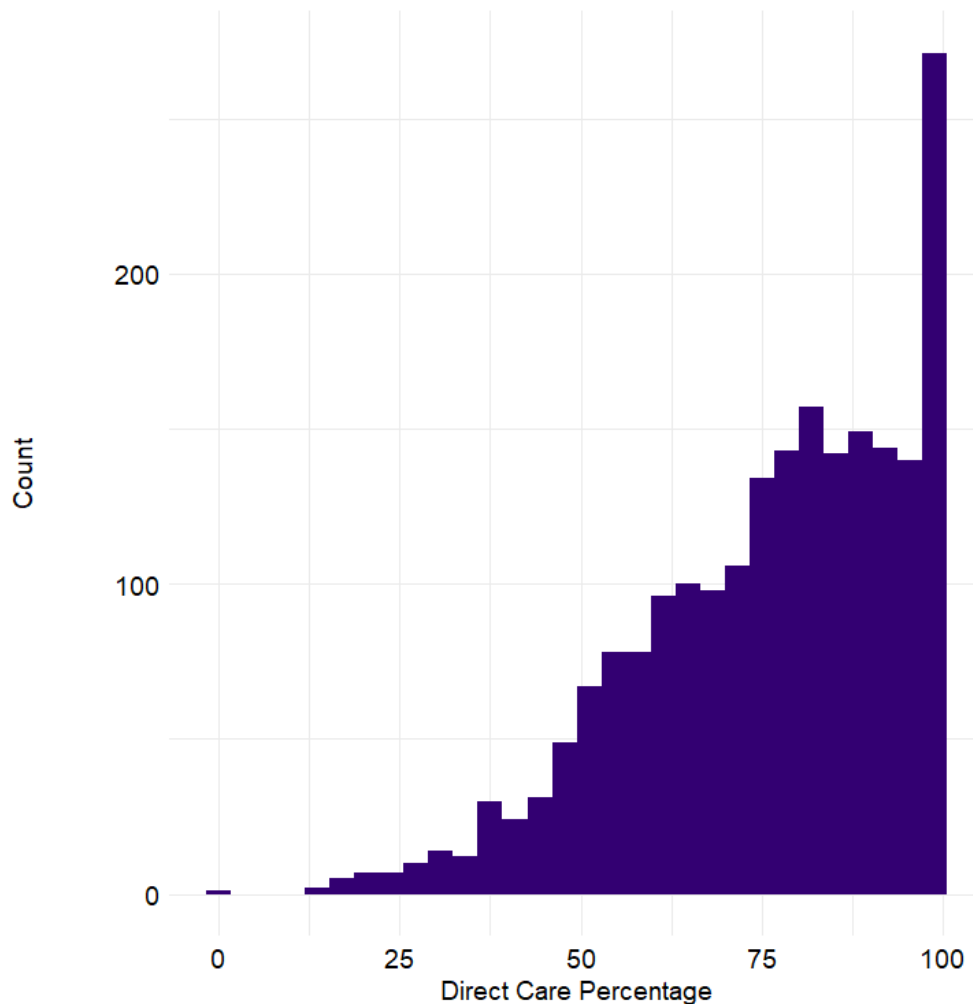


Figure E.9: Histogram showing frequency of distribution of direct care percentage

Figures E.8 and E.9 above show that the distribution of direct care percentage is bounded between 0 and 100 with a non-normal left-skewed shape and a concentration of values around 100%.

Total Time

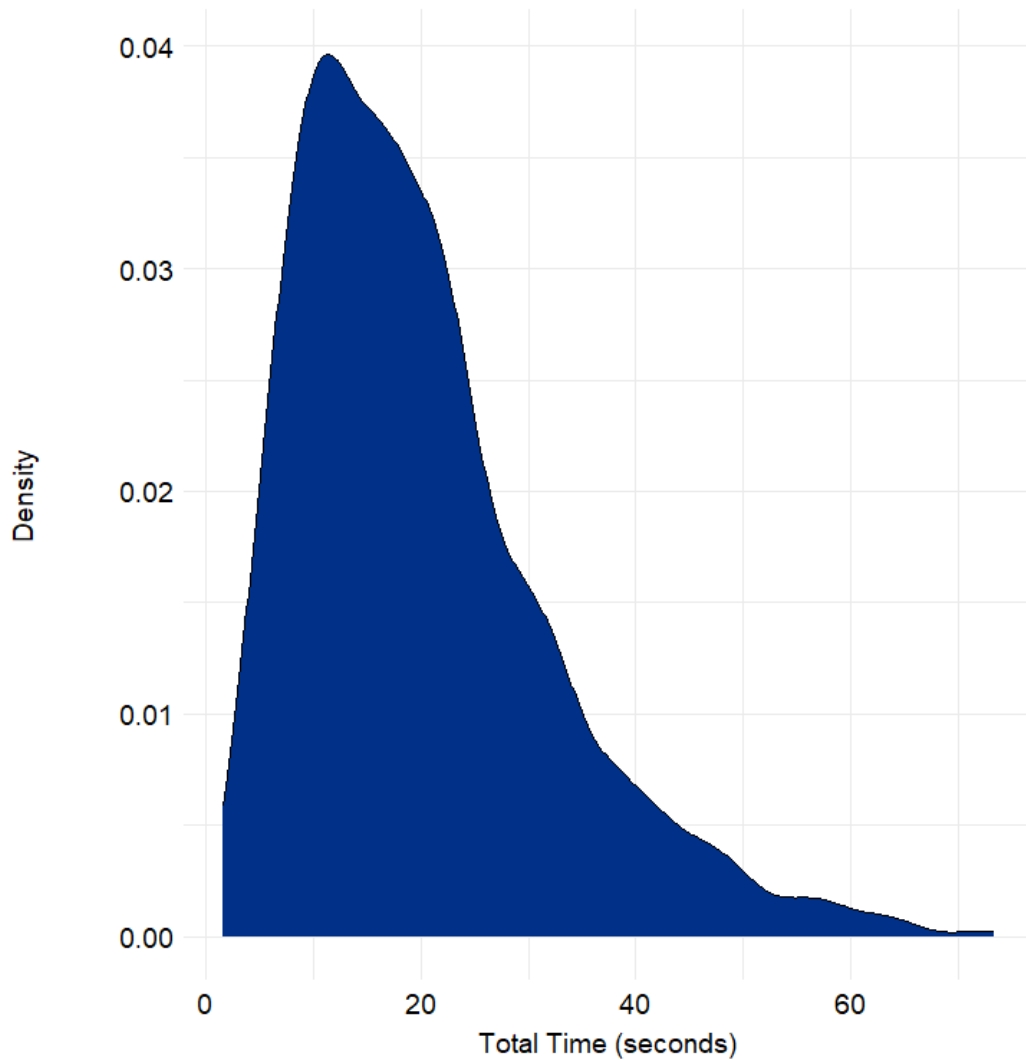


Figure E.10: Density plot showing distribution of total time

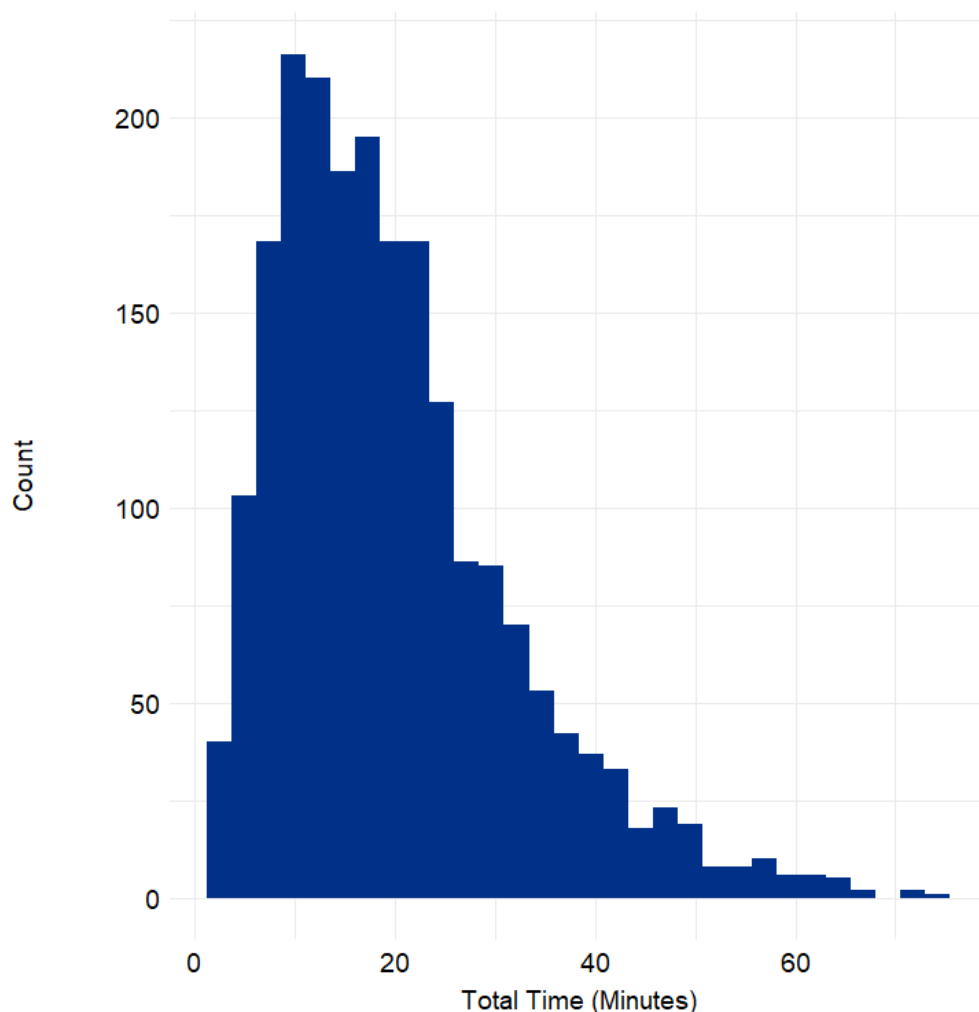


Figure E.11: Histogram showing frequency of distribution of total time

Figures E.10 and E.11 indicate that the total session time has a right-skewed distribution with a somewhat unimodal shape, where most observations fall between 10 and 30 minutes.

Model Reasoning and Residual Diagnostics

To evaluate the effect of AVT, multiple candidate models were trialled, comparing using the AIC, BIC and R-Squared coefficient.

Due to the nested data structure, we included clinician name as a random effect to account for clinician level differences and repeated measures. Though our investigation showed variation of AVT effect across clinicians (Figure E.5), it was concluded that a random-slope approach would be detrimental to the model, as this hugely increases complexity, without a satisfactory improvement in model fit or interpretability, as we wanted to focus the model on investigating effect of AVT. The

fixed effect structure was chosen based on relevance to the hypothesis and protocol, while retaining variables known/expected to influence the outcome. The random effect structure was chosen with regards to data hierarchy and removal of effects that added unnecessary complexity while not improving model fit. These effects were originally included in the maximal model and iterated out during the process of refinement, using previously stated comparison tools, as well as the intraclass correlation coefficient (ICC).

The chosen model gave a conditional R-squared of 0.91 (i.e. 91% of variance explained by model including fixed and random effects) and a marginal R-Squared of 0.41, reflecting overall strong explanatory power, with much variance being explained by clinician-level differences.

The final structure maintains the balance of interpretability and model parsimony, while still honouring the hierarchical structure of the data.

Model diagnostics were also computed within R using the DHARMA package due to the non-Gaussian nature of the model. A Q-Q plot (Figure E.12) of uniform scaled residuals and a plot of residuals versus predicted values were created, to assess model fit.

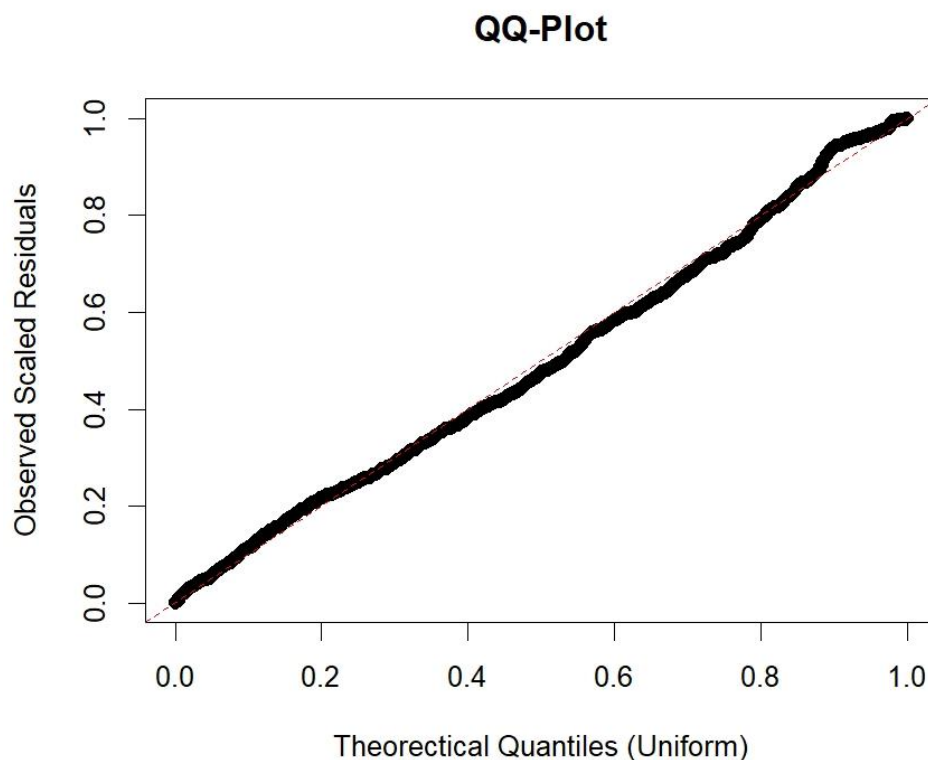


Figure E.12: QQ Plot of Uniform Residuals for Beta Model

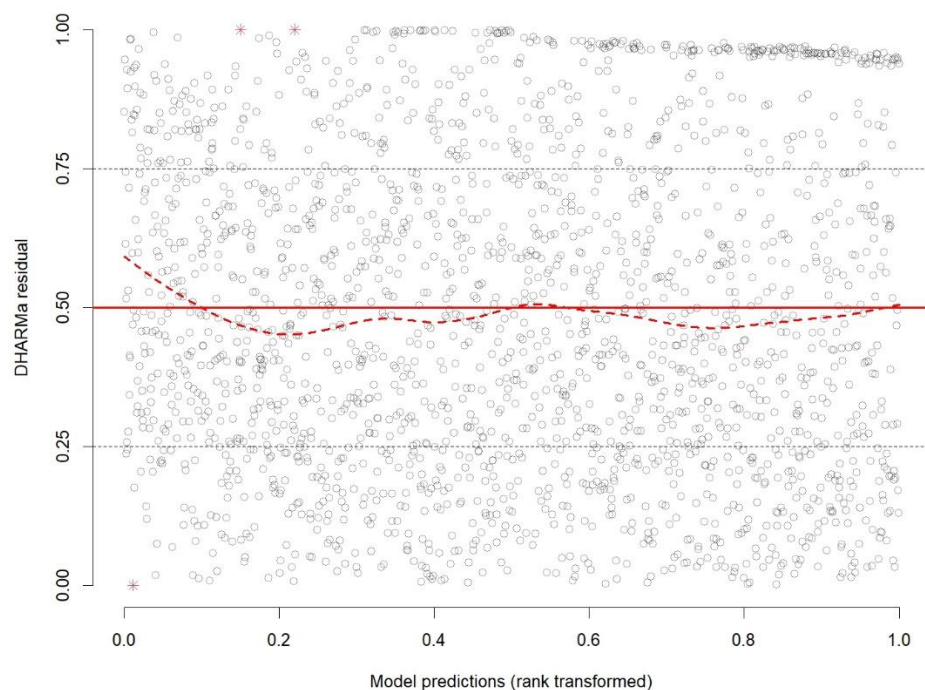


Figure E.13: Residual vs Predicted plot for Beta Model

The Q-Q plot showed no major deviations; there was a slight deviation however there was largely a close alignment between observed and expected quantiles. The residuals versus predicted values plot (Figure E.13) showed no major patterns or heteroscedasticity with residuals evenly scattered around the expected value. Although the Kolmogorov-Smirnov test and the dispersion test returned significant p-values, these minor deviations are likely influenced by the large sample size, which increases sensitivity to small effects. The zero-inflation test was non-significant, indicating the model handled the bounded nature of the outcome. These diagnostics suggest that while not perfect, the model fit is satisfactory for use in the evaluation of AVT effect on direct care.

T-test assumption testing

Direct Care

A Shapiro-Wilk test was carried out to test t-test assumptions. Despite this indicating a deviation from normality, the sufficiently large sample size ($n > 100$) meets the requirements for the Central Limit Theorem, which supports the use of a paired t-test by ensuring the sampling distribution of the mean difference is approximately normal. A histogram of the paired differences Figure E.14 below shows a moderate skew but an approximately unimodal and symmetric distribution, implying that a t-test is appropriate here.

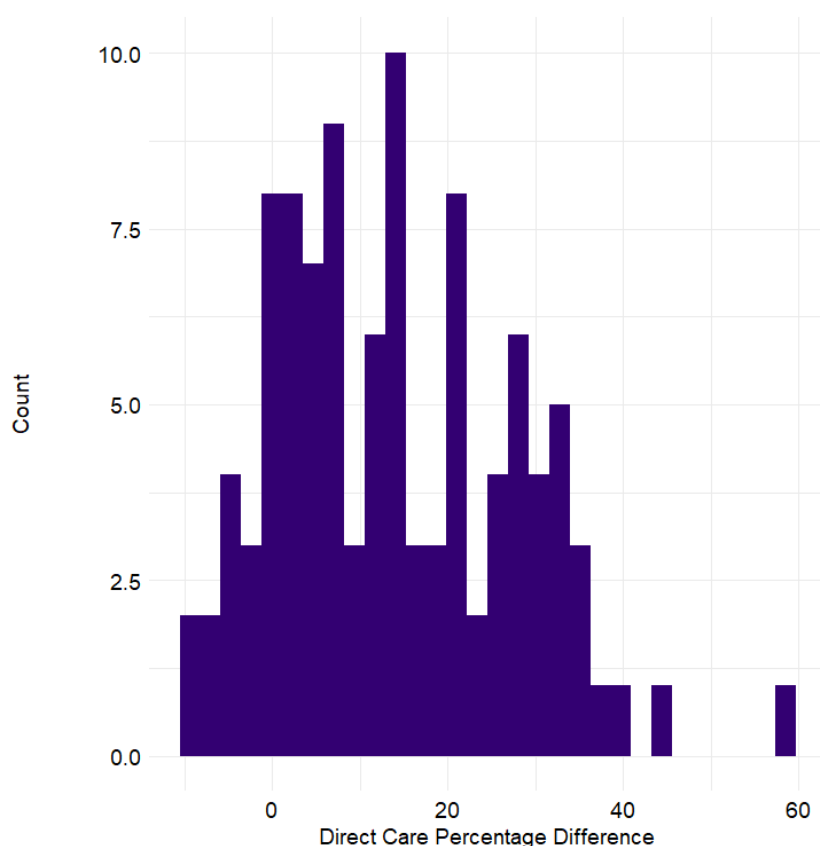


Figure E.14: Histogram of paired percentage differences in direct care time

Total Time

A Shapiro-Wilk test was carried out to test t-test assumptions. This test did indicate a deviation from normality; however, the sufficiently large sample size ($n > 100$) and the Central Limit Theorem imply that a t-test is a valid statistical method in this case. The accompanying histogram (Figure E.15 below) shows a reasonably symmetric distribution with no extreme skew, further supporting the validity of this approach.

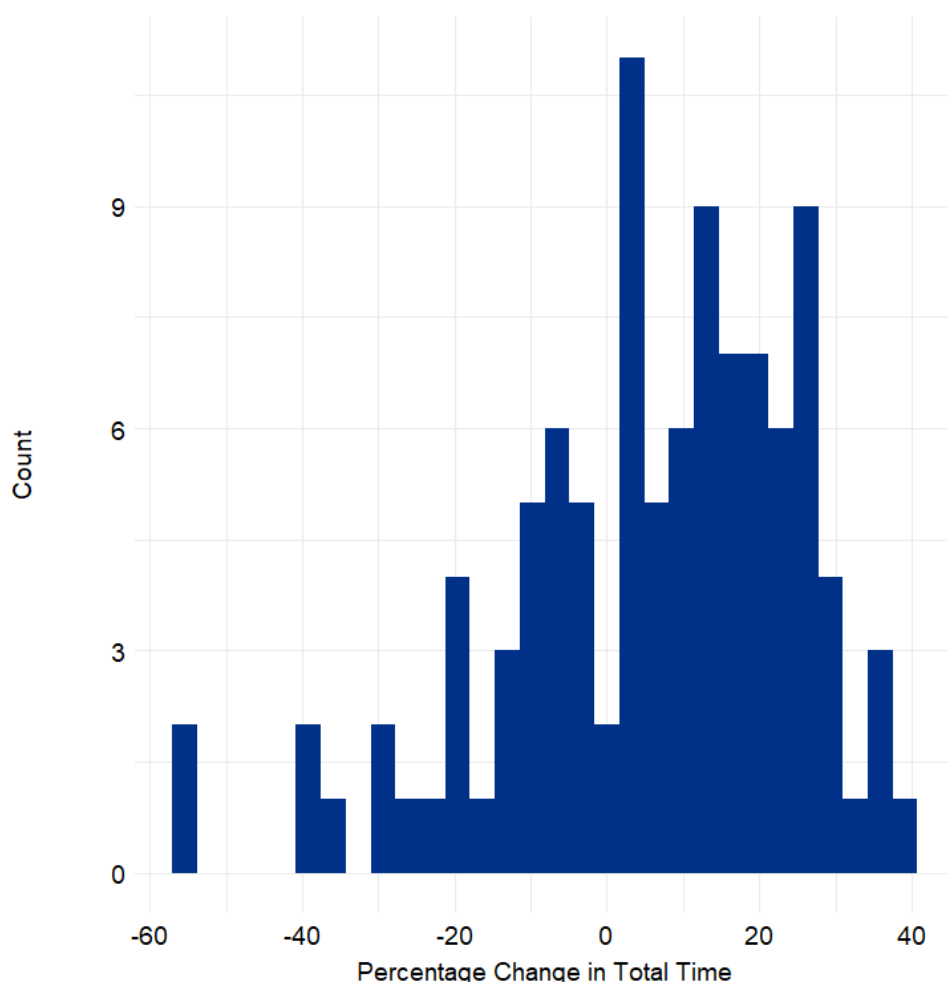


Figure E.15: Histogram of percentage changes in total time per clinician

Limitations

There are some limitations specific to the quantitative data, which need to be considered in the context of the findings:

- As shown in figure E.7, there is visible variability across the difference in session duration at baseline and AVT arms across clinicians. During the analysis, we explored multiple modelling approaches for session duration, but none provided a suitable model fit/acceptable residual behaviour, leading to our decision to use a t-test only. Though this is statistically appropriate given the data structure, it does not allow for consideration of variation of effect between clinicians.
- The beta model used had minor statistical deviations from expected behaviour of residuals, which could influence the validity of the model.

Appendix F – additional survey data analysis

Patient Survey Data

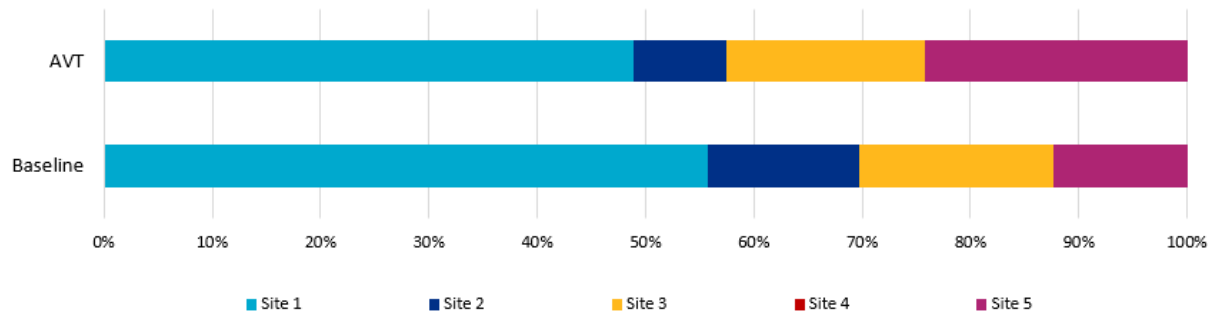


Figure F.1: Percentage patient surveys per hospital site, Baseline and AVT

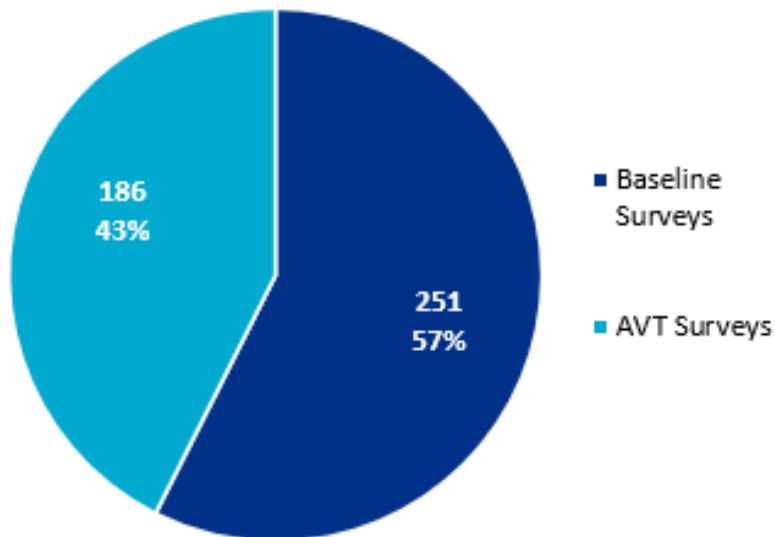


Figure F.2: Percentage patient surveys, Baseline and AVT

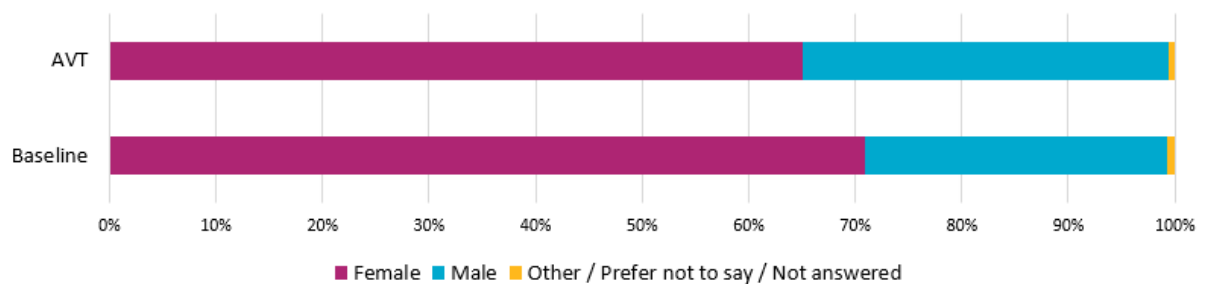


Figure F.3: Percentage patient gender at Baseline and AVT

Parent-Carer Survey Data

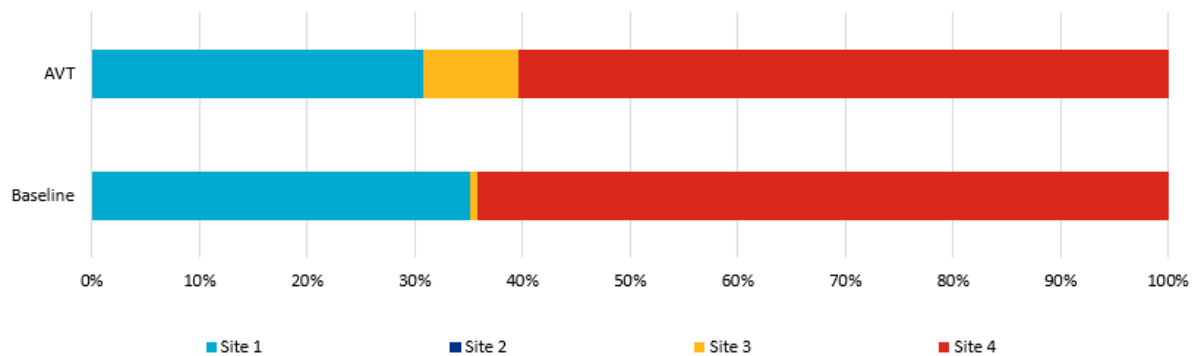


Figure F.4: Percentage of surveys completed per site at Baseline and AVT stage

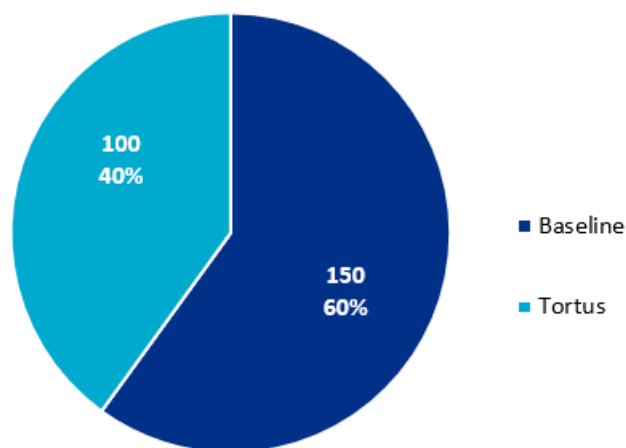


Figure F.5: Percentage surveys completed at Baseline and AVT stage

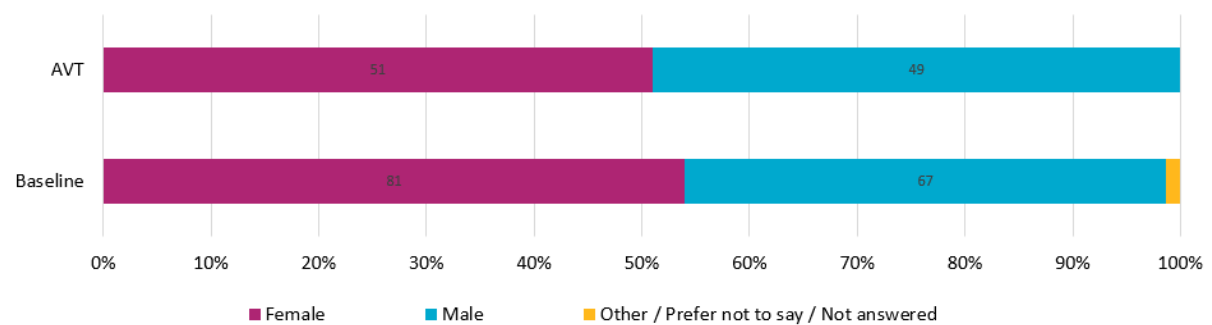


Figure F.6: Percentage of children by gender at Baseline and AVT stage.

Clinician Survey Data

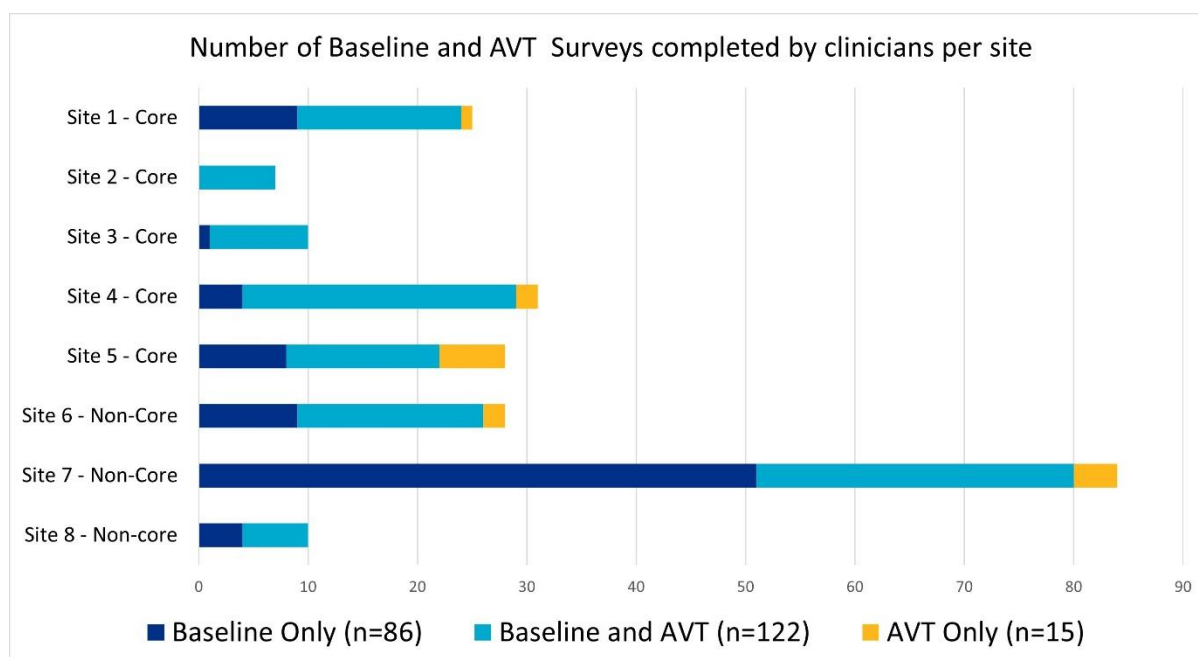


Figure F.7: Baseline and AVT Survey numbers per site. Also shown are the number of clinicians who completed both.

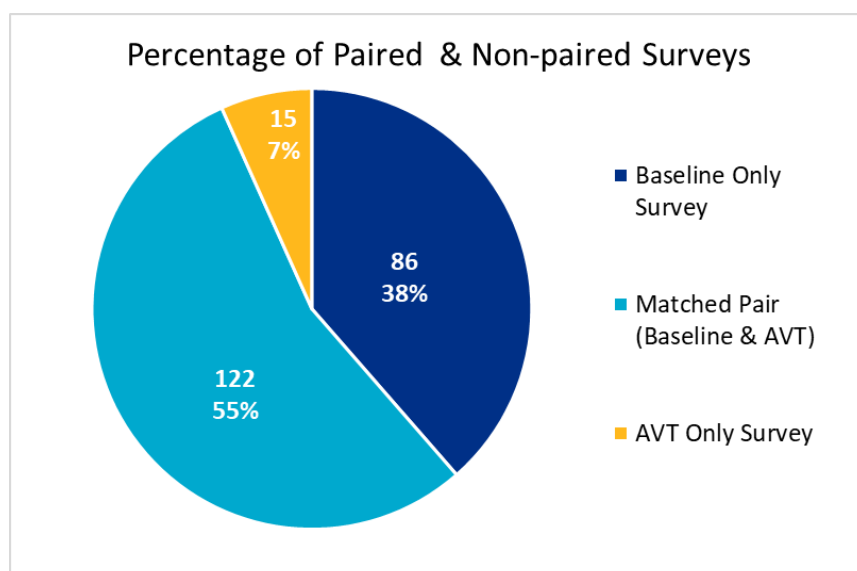


Figure F.8: Number and percentage of clinicians completing AVT experience surveys

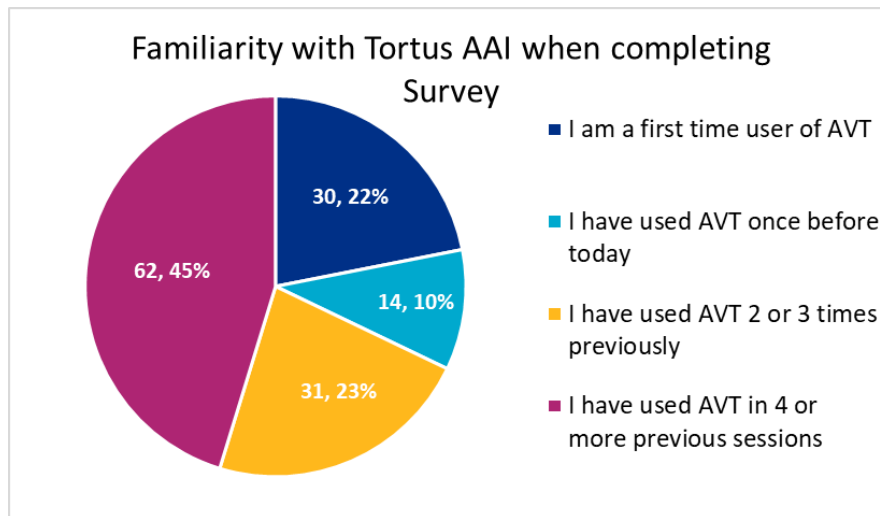


Figure F.9: Number and percentage of clinicians indicating level of familiarity with AVT

Appendix G - Sheffield Assessment Instrument for Letters

BASIC DATA

LETTER				
Code no.				
Type	new patient	follow up	referral	other
Sample	selected	random		
Status	consultant	GP	peer	self
CASE COMPLEXITY	low	average	high	

CHECKLIST

<i>Problem list</i>			
1. Is there a medical problem list?	Yes (1)	No (0)	
2. Are any obvious and significant problems omitted?	Yes (0)	No (1)	NA (0)
3. Are any irrelevant problems listed?	Yes (0)	No (1)	NA (0)
<i>History</i>			
4. Is there a record of the family's current concerns being sought or clarified?	Yes (1)	No (0)	
5. Is the documented history appropriate to the problem(s) and question(s)?	Yes (1)	No (0)	
<i>Examination</i>			
6. Is the documented examination appropriate to the problem(s) and question(s)?	Yes (1)	No (0)	
<i>Overall assessment</i>			
7. Is the current state of health or progress clearly outlined?	Yes (1)	No (0)	
8. Are the family's problems or questions addressed?	Yes (1)	No (0)	NA
9. Is/are the referring doctor's question(s) addressed?	Yes (1)	No (0)	NA
<i>Management</i>			
10. Is a clear plan of investigation or non-investigation recorded?	Yes (1)	No (0)	
11. Are the reasons for the above plan adequately justified?	Yes (1)	No (0)	NA
12. Are all known treatments, or the absence of treatment, recorded clearly?	Yes (1)	No (0)	
13. Are all doses clearly stated in formal units?	Yes (1)	No (0)	NA
14. Is adequate justification given for any changes to treatment?	Yes (1)	No (0)	NA
15. Is there an adequate record of information shared with the family?	Yes (1)	No (0)	
<i>Follow up</i>			
16. Is it clear <u>whether or not</u> hospital follow-up is planned?	Yes (1)	No (0)	
17. Is the purpose of follow-up adequately justified?	Yes (1)	No (0)	NA
<i>Clarity</i>			
18. Is there much unnecessary information?	Yes (0)	No (1)	
19. Does the structure of the letter flow logically?	Yes (1)	No (0)	
20. Are there any sentences you don't understand?	Yes (0)	No (1)	

GLOBAL RATING: (PLEASE MARK HOW MUCH YOU AGREE WITH THE STATEMENT)

"This letter clearly conveys the information I would like to have about the patient if I were the next doctor to see him or her"

1	2	3	4	5	6	7	8	9	10
Not at all									Completely

Appendix H – Net Promoter Score

Absolute NPS: what is a 'good' score?

Creators of NPS, Bain & Company, suggest a score:

- Above 0 is good,
- Above 20 is favourable,
- Above 50 is excellent, and
- Above 80 is world class.

From the absolute NPS position, any score over 0 would be considered 'good' as there are more Promoters than Detractors. Though, based on the above, it would be seen as the minimum level of progress. To be above average, a score greater than 50 is needed, so you would need to work on turning Detractors into Passives.

What is a Good Net Promoter Score? (2025 NPS Benchmark)

Appendix I - The NASA Task Load Index (NASA-TLX)

The NASA Task Load Index (NASA-TLX)

The tool is a subjective workload assessment tool developed by the Human Performance Group at NASA's Ames Research Center. It is designed to evaluate the **perceived workload** of individuals performing tasks, particularly in complex and high-demand environments such as aviation, space operations, healthcare, and human-computer interaction studies.

The NASA-TLX provides a **multidimensional rating** of workload based on **six subscales**, allowing researchers or practitioners to assess the mental and physical demands of a task from the participant's perspective.

The Six Subscales

Each dimension is rated on a **scale from low (0) to high (10)**:

1. **Mental Demand**
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching)?
2. **Physical Demand**
How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating)?
3. **Temporal Demand**
How much time pressure did you feel due to the rate or pace at which the tasks occurred?
4. **Performance**
How successful do you think you were in accomplishing the goals of the task? (Note: This is reverse-scored, where lower ratings represent better perceived performance.)
5. **Effort**
How hard did you have to work (mentally and physically) to accomplish your level of performance?
6. **Frustration Level**
How insecure, discouraged, irritated, stressed, and annoyed were you?

12. References

1. HM Treasury. New funding to kickstart delivery of two million extra NHS appointments 2024 [Available from: <https://www.gov.uk/government/news/new-funding-to-kickstart-delivery-of-two-million-extra-nhs-appointments>].
2. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. *Journal of medical Internet research*. 2020;22(7):e19274.
3. Asgari E, Kaur J, Nuredini G, Balloch J, Taylor AM, Sebire N, et al. Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review. *JMIR Med Inform*. 2024;12:e55499.
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature medicine*. 2022;28(1):31-8.
5. Hudson TJ, Albrecht M, Smith TR, Ator GA, Thompson JA, Shah T, et al. Impact of Ambient Artificial Intelligence Documentation on Cognitive Load. *Mayo Clinic Proceedings: Digital Health*. 2025;3(1):100193.
6. Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, et al. Special Topic on Burnout: Clinical Implementation of Artificial Intelligence Scribes in Healthcare: A Systematic Review. *Appl Clin Inform*. 2025.
7. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*. 2024;5(3):CAT. 23.0404.
8. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Wilson Hannay SB, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. *NEJM Catalyst Innovations in Care Delivery*. 2025;6(5):CAT. 25.0040.
9. Aldane J. AI-generated documents 'biggest bet' to improving NHS productivity, ex-director of transformation says 2025 [Available from: <https://www.globalgovernmentforum.com/ai-generated-documents-biggest-bet-to-improving-nhs-productivity-ex-director-of-transformation-says/>].
10. NHS England. NHS Long Term Workforce Plan 2024 [Available from: <https://www.england.nhs.uk/publication/nhs-long-term-workforce-plan>].
11. Balloch J, Sridharan S, Oldham G, Wray J, Gough P, Robinson R, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. *Future Healthc J*. 2024;11(3):100157.
12. Crossley J, Howe A, Newble D, Jolly B, Davies HA. Sheffield Assessment Instrument for Letters (SAIL): performance assessment using outpatient letters. *Medical education*. 2001;35(12):1115-24.
13. TimeCat. Time Motion Studies [Available from: <https://timecat.org/#statssection>].

14. Reichheld FF. The one number you need to grow. Harvard business review. 2004;82(6):133-.
15. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology. 52: Elsevier; 1988. p. 139-83.
16. Hart SG, editor NASA-task load index (NASA-TLX); 20 years later. Proceedings of the human factors and ergonomics society annual meeting; 2006: Sage publications Sage CA: Los Angeles, CA.
17. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative research in psychology. 2006;3(2):77-101.
18. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC medical research methodology. 2013;13:1-8.
19. NHS Digital. NHS workforce statistics 2025 [Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforcestatistics/february-2025>].
20. NHS England. National Cost Collection data 2023/24 2024 [Available from: <https://www.england.nhs.uk/costing-in-the-nhs/national-cost-collection/>].
21. Personal Social Services Research Unit. Unit costs of health and social care: Community-based health care staff 2023 [Available from: <https://www.pssru.ac.uk/unitcostsreport/>].
22. Department of Health & Social Care NENI, NHS Digital. Digital transformation in the NHS 2020 [Available from: <https://www.nao.org.uk/wp-content/uploads/2019/05/Digital-transformation-in-the-NHS-Summary.pdf>].
23. Sridharan S, Peters C, Readman L, Botelho B, Taylor A, Sebire N, et al. NHS TEST: not every AI is intelligent, but we need an intelligent framework to choose new technologies: Health Innovation Network South London; 2025 [Available from: <https://healthinnovationnetwork.com/resources/nhs-test-anintelligent-framework-to-choose-new-technologies/>].
24. Afzal M, French KE, Bilbrey LE, Faruki AA. Artificial Intelligence in the Clinic: Creating Harmony or Just Adding Noise? American Society of Clinical Oncology Educational Book. 2025;45(3):e481490.
25. Seth P, Carretas R, Rudzicz F. The utility and implications of ambient scribes in primary care. JMIR AI. 2024;3(1):e57673.