SW Healthcare Centre of Digital Excellence (CoDE)

AVT Testing Report 2025

About this Report

The initial focus of the South West Healthcare Centre of Digital Excellence (CoDE) has been on cutting edge artificial intelligence technologies with an overall strategy of delivering "healthier lives and improved wellbeing for all people in the South West, enabled by digital technology and data."

SW Healthcare CoDE is problem-statement driven. Problem statements have been surmised from the primary care workforce by surveys and focus groups. The statement for AVT experiments was "Inputting information is time consuming and inefficient."

Research: Research was commissioned on key challenges in primary care including but not limited to processes and pathways (administrative and clinical), based on strategic pressures and interests strictly guided by the problem statements that have been central to programme discovery work.

Experiment Design: Hypotheses were derived from the problem statement and prioritised. Experiments were then designed against these hypotheses including methods and metrics to be explored. The CoDE team were responsible for designing their programme of work packages to ensure all research questions were answered.

Experiment Set up: Experiments were set up for maximum efficiency, therefore a library of consultation recordings (audio and video) was created and used across all products.

Equipment: Hardware was sourced from the NHS, and all general practice incumbent software was installed including the most used EPRs in the South West region. One Care installed the EPRs, and a synthetic data base. One Care also advised and verified the GP space before testing commenced.

Technology source: Free versions of AVT software was sourced by the CoDE team (as this was the case across the region in primary care), who located all product information from manufacturers' websites where it was freely available. A collation of this information can be found in Appendix 2.

Participants: To complete the experiments, actors were required for the roles of patient and clinicians, prior to the validation of the results by NHS clinicians. Actors were sourced from the School of Drama at the University of the West of England.

Patient data: Synthetic dataset modules have been provided by NHS England Data Scientists to ensure data privacy and safety.

Technology scanning: Health Innovation South West undertook a "Horizon scanning" exercise of Ambient Voice Technology in Healthcare. The report identified 66 companies world-wide currently that are producing AVT products.

Executive Summary

The **South West Healthcare Centre of Digital Excellence (CoDE)**, based at the University of the West of England in Bristol, focuses on testing new and existing digital technologies used in the primary care sector in a controlled, near-to-real-world environment. CoDE includes a physical simulation environment of a General Practice, including consultation room, waiting room and back office. The CoDE facility is equipped with key software packages used in Practices. A synthetic dataset of 100,000 patient records was provided by NHS England data scientists to emulate a Practice population without compromising real patient data.

This report presents an assessment of the use of Ambient Voice Technology (AVT) in clinical consultations to review the software's efficiency and accuracy in producing a summary of the conversation. Hypotheses were derived from problem statements sourced through questionnaires to various stakeholders. Experiments were designed to test these hypotheses in CoDE using scientific methodology and metrics.

Fifty consultations were performed covering various scenarios derived from training videos and expert input (GP-patient, nurse-patient, additional person with patient and telephone consultations). Consultations were recorded and analysed with AVT and using traditional note-taking practice. Various AVT software packages were reviewed. Summaries were produced from audio files and compared with baseline data to statistically evaluate accuracy. Automated and manual comparison methods were compared. Perceptions of the consultation experience were recorded through questionnaires and video analysis.

The key findings were focused in five areas:

Experience of Consultations

Experiments assessed the fluidity of consultation, eye contact between health professional and patient, duration, and administrative time. The findings suggested that AVT improved the fluidity of consultations, increased engagement between the health professional and patient and reduced overall consultation time, through time saving on note-taking.

Accuracy of Summaries

Several factors were investigated to assess their impact on the accuracy of the summary produced using AVT. The factors tested were: background noise, misleading information, accents, colloquialisms, speech impediments, personality types, and microphone placement. It was found that errors were introduced in the presence of these interfering factors, with omissions being the most common error. Certain speech impediments, background noise and misleading information significantly impacted the accuracy of summaries. The position of the microphone had a significant impact on the accuracy of the summary. The accuracy of the summaries from a given AVT supplier was not consistent over time – many suppliers regularly update their underlying AI algorithms.

Wider Impacts

The effect of AVT on patient throughput, waiting times, and cost savings were also assessed. The findings indicated that AVT improved patient flow, and decreased waiting times. The use of AVT could also save GPs between 8 and 40 minutes per session of 15 appointments and reduces the number of times a session overruns by 90%.

Ethical considerations

Ethical considerations were identified by Health Innovation South West in a “Horizon Scan” report on AVT for the CoDE.

Data Security

The report outlines where terms of data security, there should be compliance with NHS governance and digital safety standards. Practices should utilise Information Governance, Digital Safety, and Quality toolkits, such as those provide in the report, to ensure safe implementation of new technologies.

Recommendations

All findings outlined in this report are a result of a study that took place between September 2025 and January 2025 with the understanding that this is a fast moving area of industry.

- AVT software packages selected by Practices should reach or exceed software requirements specified in hazard log provided with the report
- Placement of the microphone(s) should be a key consideration
- Training would support health professionals to handle AVT systems effectively and to recognise situations and types of errors that can occur
- Automated quality control methods would ensure on going adherence to accuracy requirement’s whichever AVT software package is selected
- Errors in summaries were produced by all AVT software packages tested - AVT-generated consultation summaries must always be checked by the health professional before saving on patient’s record.

The findings in this report are a result of testing in an offline simulated environment. This is not a decision-making document and can only be used as guidance.

This report should be read in conjunction with the **NHSE national guidance for AVT**[1].

The Health Innovation Network have published **considerations for healthcare systems**[2] Furthermore, the **NHSE national Chief Clinical Information Officer** (CCIO) published a letter dated 9 June 2025 to all CCIOs with a priority notification to ensure safe and assured adoption of AI Scribe Technology. It contained the following key points to follow:

[1] NHS England » Guidance on the use of AI-enabled ambient scribing products in health and care settings

[2] Ambient Voice Technology in the NHS – what should healthcare systems consider? - Blog

1. Do not use AVT solutions that are not compliant with NHS standards.
2. All AVT solutions that generate summarisation require, at least, MHRA Class 1 medical device status.
3. Providers need to complete a clinical safety risk assessment and data protection impact assessment (DPIA) before using these tools as part of your legal responsibilities as set out in the DCB0160.
4. Liability for using a non-compliant solutions sits with the deploying organisation (e.g. general practice or trust) or individual user.

The letter stated that it is the responsibility of all AVT suppliers to demonstrate compliance with the following requirements[3].

1.Core platform assurance requirements

a. Digital Technology Assessment Criteria (DTAC), Data Security and Protection Toolkit (DSPT,) Cyber Essentials Plus, CREST-approved pen testing
b. Data Protection Requirements as set by ICO - Local ICB / Trust governance approval including DPIA completion
c. Clinical Safety Officer(s) named and accountable
d. End-to-end encryption and GDPR compliance
e. No unsafe functionality e.g. prompt injection access
f. Appropriate NHS clinical system integration (API or FHIR/HL7 compliance and write-back capability).
g. The responsibility for translation accuracy remains with the AVT supplier.

2.Enhanced Requirements

a. **Medical Device Classification –** All AVT solutions that undertake summarisation require, at least, MHRA Class 1 medical device status. Companies must NOT extend system capabilities to produce generative diagnoses, management plans, or other medical referrals and calculations without seeking at least MHRA Class 2a approval.
b. **Data Protection –** Safeguarding Patient Information is paramount. Patient data from clinical sessions (e.g. immediate inference) should be automatically deleted unless legally or operationally required, in line with UK GDPR and DPA 2018 principles on data minimisation and storage limitation. Further guidance on this will be published shortly.
c. **System integration –** Ensure appropriate integration with your IT infrastructure, systems and workflows. For example, in most general practice and hospital settings, AVT solutions will require integration with the principal electronic record system. This will enable automated workflow (e.g. diagnostic test requesting or prescribing presented within the system being used, for clinician validation and submission).

Clinical and Operational Benefits Thresholds

a. Evidence of real-world clinical validation of benefits in the NHS care setting proposed (e.g. enhancing clinical efficiency and workflow, reducing administrative burden; improving patient care by increasing face to face time with patients; improving accuracy

[3] NHS England » Guidance on the use of AI-enabled ambient scribing products in health and care settings

of documentation; improving data quality and capture of structured data recorded in electronic patient record systems)

b. Clear economic justification and workforce impact.

Contents

Introduction

The SW Healthcare CoDE is a physical space in Bristol, set up to test new and existing digital technologies (including experimenting with combining various digital components) in a safe ‘offline’ environment and using a synthetic patient dataset, before assuring and deploying in the ‘live’ primary care environment. One Care advised on, and supported the configuration of CoDE. This enabled testing in a “close to real-world” GP environment.

The research undertaken in the Lab currently has a focus on general practice digital processes, based on a set of high-level problem statements that were identified through stakeholder engagement.

The very first experiments have had a focus on the use of ambient and generative AI in clinical consultations, and by having an approach that includes the people who have direct experience or contact in general practice and its processes, we can better assess the current state and the objective of the experiments. Overall, AVT testing provided an opportunity to determine ‘proof of concept’ for the CoDE, and an enabler for reviews and improvements in order to progress the CoDE into business as usual.

Experiments

Experiments were performed in CoDE to understand the impact of ambient voice transcription (AVT) technology. Experiments were devised and executed to evaluate AVT when it is applied to a GP consultation. The work undertaken fell into three categories:

- Experience of consultations
- Errors produced by AVT in the summary
- Wider impacts of AVT on a GP practice.

The results of the experiments addressed aspects of the following concerns.

Accuracy of summaries for patient records
Patient satisfaction and engagement
Quality of conversations (with human and environmental interference)
Time spent on administrative tasks
Patient throughput and waiting times
Cost-savings analysis
Data security and privacy concerns

Experiments were conducted using Heidi AI^{TM} system to study all the elements indicated in the table above. Further experiments were then designed to differentiate between the performance and functions that other AVT suppliers provide.

Heidi was selected as the AI Scribe (CAIS) product to test to provide baseline indicators, as it was the most prevalent in Primary Care at the time of this study (all testing took place between September 2025 and January 2025). This had been determined by a survey within primary care in the South West region. Because "reproducible results" are very important to the CoDE, a detailed account of the study is given in this section that includes the names of the product manufacturers.

A summary of the research questions addresses and our finding are found in Appendix 1.

Experimental methods

Consultations

Ten GP consultation scenarios were created from GP training videos with input from One Care. Four nurse practitioner consultations were developed from information provided by practicing nursing staff at UWE. Rather than being scripts, these scenarios provided the key points (guidance notes) for both health practitioner and patient (or additional person) to include in the consultation (one set for each party). These scenarios were validated as “typical consultations” by a practicing GP, One Care or a working nurse. Fifty consultations were performed by actors playing health professional and patients (and additional people) in the simulated GP environment. The actors included drama students and members of the School of Drama.

Each consultation was recorded and the audio and video outputs used for further analysis. Consultations were repeated with and without an AVT presence to enable a non-AI baseline to be obtained. In some cases, consultations needed to be repeated multiple times when information was not included or the conversation did not flow naturally.

The consultations covered the following situations:

32	GP – patient consultations
8	Nurse – patient consultations
6	Additional people in consultations (carer or parent)
6	Telephone consultations

Phase 2 – Production of AVT derived summaries

The audio files collected from the consultations were used to produce summaries of the consultations. Summaries produced from the original recording were used as baseline data. No relevant “noise” was included in test audio files and AVT summaries produced which were compared with the baseline data. The “noise” included irrelevant/confusing words spoken in a consultation or modified baseline recording with additional sounds added to the consultation.

Transcripts of the audio files were also prepared.

Scenarios were repeated with different actors to allow a statistical analysis of data to be performed and minimise any bias.

Production of summaries

Summaries for each consultation were produced in a first “take” in the form of written or typed notes by the health professional actor and a second one directly using the AVT software. Audio files of the consultation were also collected. Summaries produced from the original recording, with no noise or additional factors introduced, provided baseline data. These could then be compared with modified AVT summaries e.g. ones with addition of irrelevant/confusing words spoken in the consultation or with background noise, to establish the effect of these interfering factors.

Scenarios were repeated with different actors to allow a statistical analysis of data to be performed and minimise any bias. People of different genders and ethnicities acted as both health professionals and patients.

A practicing GP generated notes from the video/audio files of the consultation which were compared with those generated by the actors (and the AVT) to ensure that the health professional actors produced notes that were of a similar detail to those produced by working professionals.

Video and questionnaires

To establish each parties' perception of the consultation experience, questionnaires were completed by health professional and patient actors following each consultation. The following were considered:

- Eye contact
- Fluidity
- Time taking notes
- Duration

The data were recorded either as a score or as perceived times. Data from the AVT consultations were compared with the equivalent where traditional note-taking was used. Nonparametric statistical analysis was performed.

Extensive analysis of the video recordings of the consultations was performed to assess the interaction between the health professional and patient. The following parameters were measured:

- Duration of consultation
- Post consultation admin time
- Total consultation time
- Eye contact – GP and Patient
- Joint eye contact
- Median dwell – length of each instance of eye contact - seconds
- Frequency/min – how many eye contacts per minute

Data was recorded in minutes and analysed using nonparametric statistics.

To establish the accuracy of an AVT-generated summary of a consultation, the following manual process was completed. The transcript of the consultation was manually corrected relative to the audio files. A table was generated with one column including each of the points from the corrected transcript and second of the key points from the summary. Finally, the points from the two columns were compared to identify differences.

Similarly, to manually compare a base-line summary with a modified summary (i.e. summary of the consultation with interfering factors included), a similar table was generated with columns for key points from the two summaries.

The differences were classified as errors under the following three categories:

- Inaccuracies – information in the modified summary from the audio file that was absent in the baseline summary.

- Omissions – information that is absent in the modified summary that was present in the baseline summary or in the audio file.
- Hallucinations – information in the modified summary that is not present in the baseline summary or the audio files.

The analysis of the summaries was performed by three people and a consensus figure for the 3 types of error was produced. Numbers of errors were visualized using Box and Whisker plots (graphs) and analysed using non-parametric statistics. The error classification performed by the Delivery Team was validated through checking by a working GPs who independently reviewed the differences between the modified and baseline summaries. There was almost complete correlation between the Team's analysis and the GP assessment.

The manual method for comparison of AI summary is considered the gold standard in the academic literature but it is highly time-consuming to perform. In order to seek more efficient methods of assessment particularly in the case where a quality assurance process is considered appropriate, an automated approach was also investigated.

Advanced analytical techniques which are used in the broader AI field were investigated and adapted for this AVT application. Initially, the Damerau-Levenshtein[4] method and library were used. In this technique the number of characters that need to be inserted, removed, replaced or transposed with an adjacent character to make the modified transcript the same as the baseline are determined. The ratio of the Damerau-Levenshtein distance/number of characters in the baseline document indicates how different the modified version is compared with the baseline. Utilisation of this method required creating of software to implement the Damerau-Levenshtein algorithm and to automate running of the tests.

A second technique to evaluate differences between sentences of the modified summary and the baseline summary was developed based on an advanced method of natural language processing, called for sBERT (sentence-Bidirectional Encoder Representations from Transformers). sBERT establishes the semantic similarity between phrases/sentences and therefore provides a more meaningful comparison two text files. A software script was written using the sBERT framework to compare the AVT summaries and to provide a method to automate comparisons of the summary.

These methods of analysing the data employed a mixture of manual and computer evaluation which allow a high degree of confidence in the results obtained. The results give a good reflection of how ACT systems would perform in real world situations.

[4] Damerau–Levenshtein distance - Wikipedia

Results

Experience of consultation

In this group of experiments, consultations were run, based on the scenarios described above. Experiments were repeated with AVT and with traditional, typed or written note-taking to understand the effect on the experience of using AVT against a baseline without AVT. Consultations were run with actors playing the GP and the patient, plus also additional person (carer/parent) where appropriate.

The parameters measured were:

- Fluidity of the consultation
- Duration of consultation
- Post consultation admin time
- Eye contact – GP and Patient
- Joint eye contact
- Median dwell – length of each instance of eye contact - seconds
- Frequency/min – how many eye contacts per minute

Statistical tests were performed to assess impacts of AVT compared with traditional consultation with typed or hand-written note-taking.

Findings

The use of AVT significantly improved the health professional and patient experience of the consultation.

- Consultation perceived to be more fluid, both for the health professional and patient
- Greater health professional engagement with the patient
- Reduced overall consultation time, with health professionals spending less time on admin activities

The majority of these results are platform agnostic. Certainly, it would not be apparent to the patient which software package is being used. The only variant is the length of the summaries produced by different suppliers (and settings/templates) which impact on the health professional checking time and thus the overall consultation.

Accuracy of summaries

Experiments were designed to assess the additional errors caused by a variety of human or environmental factors impacting on the consultation. This either required re-running of consultations or additions to the audio files of baseline consultation before summarising using the AVT system. The resulting modified summaries were compared with the baseline summary to evaluate accuracy of the modified summary and introduction of new errors. The accuracy of the summary produced by AVT technology was reviewed in the following situations:

- Misleading information
- Misleading medical information
- Background noise
- Different accents
- Understanding colloquialism
- Speech impediments
- Personality
- Microphones
- Settings and platforms

Misleading information

In this situation, additional non-medical information was added to various consultations and compared with baseline. To ensure the effect of the additional statements could be clearly established each baseline audio file was spliced with another audio file including additional statements and then audio processing techniques used to cover the “cuts”. Thus the new audio file was identical apart from the additional information. The information comprised of the following themes:

Scenario	Added information
Diarrhoea	Patient very embarrassed, wife suggested to visit.
Headache	Discuss tennis club and summer league.
Prostate	Patient embarrassed, long pauses and NDA discussion.
Skin rash	Long discussion of child care and friends.
Memory Loss	Long chat of non-relevant sport injury.

Tests were performed to assess the effect of the quantity of additional information by only giving half the amount of additional information in half of the scenarios.

In all cases there was a statistically significant loss of information in the modified summary. There was no significant difference in the accuracy or number of errors in the amount of misleading information given, but the more additional information given did tend to give rise to a higher error rate.

All cases where additional information was included had at least 2 additional errors with a maximum of 7 errors. Overall there were twice as many omissions as inclusions with only 1 hallucination noted. The median error rate was 16.7% for the number of omissions and 7.2% for inclusions.

Misleading medical information

In this situation additional non-relevant medical information was added to different consultations and compared with baseline. A set number of points in the baseline were assessed. The additional information comprised of the following:

Scenario	Added information
Diarrhoea	Patient reemphasises bowel cancer and talks of IBS.
Headache	Patient reemphasises brain tumours.
Prostate	Patient reemphasises UTI possibility.
Skin rash	Patient additionally talks about chicken pox.
Memory Loss	Patient reemphasises Alzheimer's.

An assessment of impact of the quantity of additional information was performed by generating further audio files with only half the amount of additional information.

In all cases there was a statistically significant loss of information in the modified summary. There was no significant difference in the accuracy or number of errors in the amount of misleading information given

All cases had at least 1 error with a maximum of 5 errors. Overall there were 30% more omissions as inaccuracies with no hallucinations noted. The median error rate was 9.1% for the number of omissions and 4.7% for inaccuracies.

Background noise

To assess the effect of extraneous noise on the consultation, software was written to incorporate various sounds at a variety of volumes, relative to the voices of health professional and patient (additional person) in the consultation. The number of errors and accuracy of the summaries were measured both manually and using the automated comparison, as described above. Four different sounds were used:

- Baby
- Toddler
- Construction
- Heavy rain

These were mixed with the audio files at 4 different volumes: -10, -5, 0, +5 dB

Results showed that errors were introduced at all levels of sound, but unsurprisingly the number of errors introduced into the summaries increased as the volume increased. On average this rose from 1.8 errors at the lowest volumes of interfering sound to an average of 8.75 errors at the highest volumes. In the worst case up to 24 errors were introduced.

Omissions were the most common error, 20 times more common than inaccuracies and 8 times more common than hallucinations. Across all the scenarios, at the lowest volume only 1 inclusion and 2 hallucinations were observed compared with 50 omissions. This compared with 167 omissions, 7 inaccuracies and 20 hallucinations at the highest volume of sound.

Inaccuracies, as assessed by sBERT, were introduced at all volumes which increased as the volume increased. A correlation between the sBERT score and the number of errors was found.

The sounds which gave the greatest number of omissions and hallucinations were heavy rain and a toddler. The baby crying gave a higher proportion of inaccuracies.

Accents

The baseline consultation audio files were modified to emulate accents in a controlled manner. This was technically extremely challenging. Some open source cloning software was available, e.g. Applio, but these did not work well, with too much of original phonetic features, from original recording, bleeding through into the cloned speech. In addition, the text to speech raised issues with the transcript being derived from an original recording as a verbatim transcript (including all the um's and err's, repeated words etc) is needed. Finally, a solution was achieved that effectively "stitched" the transcription and the additional words (once identified) together. It was also important to consider a methodology that could be easily scalable if the advantages of the automated approach were to be realised. A neural network programme was then used to adapt the text to the required accent. (There are commercial packages, with libraries, that support this, but these were too expensive.) This required about an hour of audio recording which presented difficulty in finding recordings that could be legitimately used (copyright issues). Some legally-unencumbered voice clones were found but ideally it would be best to explicitly recruit people with strong accents and consent.

Findings indicated that board accents caused a loss of information in all scenarios tested. There was not a significant difference between different accents tested (Good English, Chinese, Indian, Scottish) whether this was attributed to the GP or the patient.

Understanding colloquialism

Various common phrases were inserted in to baseline audio files using the splicing technique described above. The resultant summary generated by the modified audio files was assessed for errors relative to the baseline summary. Phrases introduced were:

- Water tablets
- Ticker
- Febrile
- Bugs in urine
- CABG

With the exception of the introduction of the term "febrile" all other situations generated errors in the modified summary. There were no hallucinations and overall twice as many omissions than inaccuracies. On average the error rate was 5.9% with the highest error rate being 22%.

Speech impediments

Recordings of people with speech impediments (with copyright consent) were played in to the AVT systems and the number of correctly identified words in the baseline and modified baseline identified. (There were ethical issues associated with attempting to clone voices with speech impediments. In the audio files, some of the voices were quite challenging to understand. The following speech defects were assessed:

- Phonological impairment
- Vowel disorder
- Childhood apraxia of speech
- Articulation disorder
- Cleft palate

Patients with a cleft palette had the least number of differences between the baseline and modified scripts. This was followed by patients with vowel disorder where a wider spread of difference was observed. The other speech impediments all showed a statistically significant number of differences in the modified script compared with the baseline script.

Personality

Actors playing 5 different personality types underwent consultations involving 4 scenarios. The summaries produced by AVT were compared with baseline summaries of the same scenarios. The 5 personality types investigated were:

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

The most common error reported was that of omission being 4 to 5 times more frequent than inclusions or hallucinations. The character trait of agreeableness gave the fewest errors, no inclusions or hallucinations and only a total of two omissions. In contrast, the personality type of extraversion gave significantly higher numbers of omissions, being twice as high as any other personality type.

Microphones

The following microphones were tested.

FiFine K668
Logitech C920
Konftel Ego
Neat Skyline

Baseline measurements were made with the microphones on the desk and then audio files of consultations played into the microphones at different positions from the desk.

- Position 1 - 0.5 m
- Position 2 - 2.0 m
- Position 3 - 4.5 m behind an obstacle

The average number of total errors rose significantly as the distance to the microphone increased. The most common error being an omission. The number of errors ranged from an average of 3.77 omissions at position 1 to an average of 22 omissions at position 3 (12.1 at position 2). At position 1 85% of consultations had at least 1 omission. At position 2, 95% of consultations had at least 1 omission and 100% of consultations at position 3.

At position 1, only 15% of consultations gave rise to 1 or 2 hallucinations which doubled to 30% of consultations when the microphone was in position 2. At position 3 there were no hallucinations or inaccuracies noted in any of the consultations. This was the quality of the conversation at position 3 was so poor that the modified summary had lost all information with 100% of the points of information in the baseline summary being lost in 80% of the consultations and at least 70% loss of information in the others.

Configuration of settings and platforms

Heidi (free) configurations

Audio transcripts were prepared from ten different scenarios from audio files of the consultation. Exactly the same audio files were run through different settings of Heidi AI. Resulting summaries were manually compared with the original audio transcript and the numbers and types of errors counted.

- Heidi has 4 different settings:
  - Brief
  - Goldilocks
  - Detailed
  - Super detailed

- Each of which has 2 different modes:
  - Left hemisphere: “Fast and unimaginative. Can lack detail. Good for simple sessions.”
  - Right hemisphere: “Thoughtful, but slow. Can infer meaning. Good for complex sessions.”

The month that testing was performed on the AVT platform was noted as the suppliers update the Large Language Model in a way that is not evident to the user.

Heidi settings tested:

- Brief/Left (December)
- Detailed/Right (December)
- Goldilocks/Left (December)
- Goldilocks/Left (Original - September)
- Goldilocks/Right (December)

Comparisons were made between summaries originally generated by Goldilocks/Left setting in September and summaries generated by other settings in December. It was noted that there were statistically significant differences in the number of errors generated using different setting to generate the summary.

All the settings generated errors (between 0 and 6 errors), with omissions being the most prevalent. Compared with the summaries generated in September, three of the four settings had significantly higher omissions, only the Detailed/Right setting being comparable. There was a significant difference in the number of omissions in the summaries generated in September and December using the Goldilocks/Left setting. The later summaries generated in December had a greater number of errors.

Platform analysis

AVT platforms that offered a free trial were reviewed using the suggested setting where there was a choice. Details and features of the different AVT platforms are given in appendix 2.

Audio transcripts were prepared from ten different scenarios from audio files of the consultation (as above). The same audio files were run through different AVT platforms using their default settings. Resulting summaries were manually compared with the original audio transcript and the numbers and types of errors counted.

The AVT platforms tested were:

- Heidi AI
- Tortus
- Kiwipen
- Nabla
- Lyrebird
- CortiAssist
- ConsultNote

The was a wide variation in the number of errors generated by the different AVT platforms (between 0 and 11). The median number of errors ranged from 1 to 6 with ConsultNote, CortiAssist and Kiwipen generating significantly number of errors. Omissions were the most prevalent of the errors generated, often significantly greater than inaccuracies or hallucinations. Lyrebird, Nabla, Tortus and Heidi G/L had similar performance.

These results are based on the free versions that are available from manufacturers, with corresponding information freely accessible via their websites. In future case studies, all manufacturers will be contacted and offered the opportunity to engage with the SW Healthcare CoDE directly.

Technology scanning

In a report produced by Health Innovation South West 66 companies, world-wide, were identified that market products with a focus on AVT products. Over 50% of these being US based companies with 5 companies located in the UK. The report then gave a more detail review of 18 companies; 11 Start-ups, 3 electronic patient record providers and 3 large enterprises.

Overall Findings

There are many situations in which errors are introduced into the summary generated by the AVT platform. Based on the number of errors generated in our experiments the following table has been produced to show the relative impact of different factors investigated have on the accuracy of the summary produced by an AVT platform. The more stars, the greater the impact and more care required in controlling the factor.

★★★★★★	Placement of microphone
★★★★★	Some speech impediments
★★★	Loud background noise
★★★	Construction work
★★	Colloquialism
★★	Misleading terms - general
★★	baby crying
★	Misleading terms - medical
★	Accents
★	Extraversion personality

Caveat: This ranking is based on experiments performed in the CoDE and may be revised as more experiments are performed.

Baseline measurements were important in determining the effect of confounding factors.

Development of a hazard log

Based on the results from the experiments a hazard log has been created. It identifies "Hazards" based on the experimental evidence. From an estimate of the likelihood and impact a risk assessment is given with suggested software requirements and control measures.

It should be noted that the hazard log is a document produced from experiments conducted in the CoDE, and clinicians and purchasers will need to define their own requirements. The evidence from experiments has informed the "suggested software requirements" including the values/ranges presented.

In some sections, it is proposed that standard audio files could be provided (from the library we have created) to support AVT system set up and potentially, to be used as a quality control process.

The hazard log is a separate document supplied alongside this report. The current version is Hazard log AVT v5.xlsx

Hazard log AVT v5
(1).xlsx

Wider Impact

The impact that AVT platforms could have on the day to day running of GP Practices, was gauged by developing statistical models based on the results of the experiments performed in CoDE, input from GPs who have used AVT in their consultation and from Practice Managers at One Care.

The statistical model assessed 50 GP sessions, each with 15 appointment slots of 15 minutes. Variability of the consultation was normally distributed giving consultations between 11 and 20 minutes. The model included the probability of the GP being interrupted at some point during the session for between 2 and 8 minutes. The use of AI saved between 1 and 5 mins per consultation.

Impact on administration time

Based on the experiments performed in CoDE, there was a statistically significant reduction in the total consultation time of up to 5 minutes when AI was used in the consultation.

The statistical model suggested that the use of AVT platforms by GP in consultations would save a GP between 8 and 40 minutes in each session. The number of overrunning sessions would be reduced by 90%.

Impact on patient flow

The model assessed how long patients would be waiting to see the GP, assuming they arrive between 2 and 8 minutes prior to the allocated appoint time.

Time patients waited in waiting area

Without AVT, waiting times increased during the session period with patients waiting up to 10 minutes at the beginning of the session to between 10 and 50 minutes at the end of the session.

The use of AI in the consultation reduced the average waiting time over a session from between 10 and 30 minutes to less than 10 minutes. This also had the effect of increasing the number of patients waiting 10 minutes or less from 7% to 89%.

Number of patients in the waiting room

Here the model assumed 5 GPs working in one session of 15 consultations, each of 15 minutes. The other parameters were the same as used in the other studies.

For the first 45 minutes of a session there were up to 6 patients waiting for their appointment in the waiting area. The number of patients waiting increased as the session progressed with between 7 and 12 patients waiting for their appointment. The use of AVT by the 5 GPs resulted in only one, sometimes 2 patients waiting for their appointment in the waiting area.

Caveat: The model assumed all GPs work at the same pace and there were no significant external factors that prevented GPs from doing their consultations. Results presented are probably best-case scenarios but give an indication of the scale of the potential impact of using AVT platforms in a GP practice.

Cost

There is significant cost implication in installing and running AVT systems in a GP practice, and this should be taken into account following the use of a free trial.

Ethical considerations

Health Innovation South West identified the following ethical issues that would need to be considered in the adoption of AVT systems.

Informed consent

One of the central ethical concerns is ensuring that patients and healthcare professionals provide informed consent before any data is captured or processed. Patients need to be fully aware that their voice interactions are being recorded and used, and they must have the opportunity to opt-out.

Data ownership and use

- The data collected by AVT (such as recordings, transcriptions, or even analyses of conversations) may belong to various stakeholders - patients, healthcare professionals and the innovator. This raises questions about who has the right to access, use, or share the data and how long the data is stored for and where (i.e. inside or outside of the UK).
- There may be concerns about whether data collected for one purpose (e.g., clinical documentation) could be used for other purposes, such as research or commercial purposes, without patient knowledge or consent.
-

Bias and inequality

- AVT can exhibit biases based on factors like accent, dialect, gender, speech impairments, or language proficiency. If the technology is not sufficiently trained or calibrated, it may misinterpret or fail to understand certain individuals, leading to inaccurate or incomplete data.
- There is also the issue of equity in access to these technologies. Not all healthcare professionals or patients may have equal access to AVT, which could exacerbate existing health disparities.
- If the technology incorporates AI that uses historical or biased data, the technology could perpetuate existing inequalities, leading to unjust decisions regarding treatment or care.

Accountability and transparency

- If an error occurs due to the AVT (such as incorrect transcriptions, missed commands, or security breaches), it can be challenging to determine who is accountable: the healthcare professional, the innovator, or another party. Clear guidelines on accountability and transparency are therefore necessary.
- If AVT integrates AI for clinical decision-making there are ethical concerns about the accuracy, fairness, and transparency of these algorithms. Healthcare organisations must ensure that AI is used responsibly, and clinical decisions remain within the hands of qualified professionals.

Patient trust

Continuous voice monitoring can create a sense of surveillance among patients and healthcare professionals, which could hinder open and honest communication. Patients may withhold critical information if they feel their privacy is being compromised, impacting the quality of care.

Environmental impact

AVT systems rely on large-scale data processing, often involving cloud-based servers and AI models, which require significant computational power and energy.

- Training and deploying large language models for AVT contributes significantly to carbon emissions.
- Continuous operation of AVT in healthcare settings, such as real-time transcription and analysis, adds to energy demands.
- Balancing the benefits of AVT against its carbon footprint is essential, particularly as healthcare facilities work to align with sustainability goals.

Data security

In BNSSG (Bristol, North Somerset and South Gloucestershire) organisations that support General Practices (see definitions below) have developed an *Emerging Tech Process* to support the safe implementation of new technologies. As part of this support, Practices have access to an *Information Governance, Digital Safety and Quality toolkit*. This toolkit helps to ascertain whether a digital tool will be safe and useful once implemented within primary care.

The completed toolkits (see Appendix 3) allow General Practices to have access to a background check on the supplier, requires the supplier to provide documentation that evidences that they have safety controls in place and asks important questions about issues such as the location of servers if patient data is handled or stored. Practices are able to nominate suppliers to be progressed through the emerging tech process, to ensure they can safely use their products.

One Care also supports Practices by providing a ‘discharge summary’ document which explains what happens after the emerging tech review process ends (Appendix 4). The discharge summary explains how duties and responsibilities are handed back to General Practice and the likely next steps involved for both parties. This ensures that Practice staff are fully aware of their ethical and legal responsibilities, supporting practices in their due diligence and enabling them to make the most informed decisions based on the gathered information.

AVT technology suppliers are regularly progressed through the emerging tech process to ensure that this technology can be safely accessed by Practices.

Appendices

Appendix 1 - Research questions addressed

Appendix 2 – Summary of AVT platforms: all information was located from manufacturer websites

The following are supplied by One Care and remain their Intellectual Property:

Appendix 3 - One care Governance, Digital Safety and Quality Toolkit

Appendix 4 - One Care Tech review discharge summary

Appendix 1 - Research questions addressed

Research question	Finding
• Does the summary information recorded using transcription software included all key clinically important elements of the consultation?	Not in all situations where additional information may confuse the software or if there is external factor influencing the quality of the recording
• Is the key information recorded and irrelevant information omitted?	In most circumstance key information is recorded accurately but there are situations where information id omitted
• Can distractions or misleading statements (for example self-diagnosis by the patient/carer) be ignored in the summary (or referred to appropriately)?	Yes - both general and medically non-relevant content in the conversation can be incorrectly included in the summary
• What are the limitations relating to background noise both general, e.g. building work and generated by patient and/or accompanying person?	Background noise above -10 dB will cause the inclusion of errors in the summary. Different type of noise and increasing volume of noise can have a worse effect.
• Is the summary information complete when presented with different accents?	There are errors introduced with accents. More work required to evaluate which accents give rise to more errors.
• Are alternative meanings (in terms of sense and reference) of words/clauses correctly presented?	No - use of unrelated medical and non-medical terms can introduce errors into the summary
• Is the summary information complete when patient has a speech impediment?	No - some speech impediments cause the loss of all information in the summary.
• Is the information recorded consistent (complete) even if presented in different forms (from people with different personality traits)?	Some personality traits induce more errors into the summary
• Is the hardware set up critical to accurate recording?	Yes - the quality and placement of the microphone is critical in minimising errors
• How does the use of ambient AI for real-time voice recognition and transcription impact the accuracy of patient records?	In certain situations, errors could be transferred from summaries into patient records if the summary is not vetted by the GP
• How does the use of ambient AI affect the quality and depth of conversations between clinicians and patients?	In general, the use of AI enhances the quality of the clinician-patient interaction
• How does ambient AI impact the time spent by clinicians on administrative tasks?	In general, the use of AI reduces time spent on admin tasks
• What effect does ambient AI have on patient throughput and waiting times in GP practices?	The use of AI reduces patient waiting times in the surgery

Appendix 2 - Collation of details and features of AVT platforms tested (all information available from their websites)

Name	Web address	Free trial?	Notes	Pricing	Claims NHS Compatible	Claims alrady in NHS use	T&C's Issues Noted	Data Processed & Stored in UK	Claims Cyber Essentials	Claims Other UK Standards?
Heidi	https://www.heidihealth.com/	Yes	Basic package is free, includes clinical notes but not "pro actions"	"Pro" is £33pm per user (individual users). "Together" is £50pm per user (allows MFA and other things)	Yes	Yes	None	Yes	Yes	DCB0129, DTAC, DSPT, GDPR
Tortus	https://tortus.ai/	Yes	10 hours a month free	£79pm for single account.	Yes	Yes	Clinician use only. No benchmarking allowed	Yes	Yes	DCB0129, DTAC, DSPT, DPIA, GDPR
Kiwipen	https://www.kiwipen.com/	Yes	Limit of 40 consultations	£30 pm or £300/yr for a single account. Offers "flexible pricing" for teams such as GP surgeries etc.	Yes	No	Clinician use only.	Unclear	Yes	DSPT, GDPR
Nabla	https://www.nabla.com/	Yes	30-day trial	$119 pm for a single account. Flexibility for larger teams isn't explicitly detailed (but expected).	No	No	Clinican use only.	No - USA	No	GDPR
Corti Assistant	https://assistant.corti.ai/	Yes	14-day trial	$99 pm for a single account. Offers "flexible pricing" for teams such as GP surgeries etc.	Yes	Yes	None	No - Option EU or USA	Yes	GDPR, DCB0129, DSPT
Lyrebird Health	https://www.lyrebirdhealth.com/uk	Yes	14-day trial	£59pm for single account. Offers flexible pricing for larger teams.	No	Yes	None	No - Australia	No	GDPR
ConsultNote	https://www.consultnote.ai/	Yes	14-day trial. Possibly Australia only	"Introductory pricing" specified only: $190 AU pm.	No	No	None	No - USA (and possibly Australia) (uses OpenAI and Google)	No	None

Name	Time summaries are stored	Settings Flexibility	Generates referal letters	Provides diferential diagnosis	Generates Clincal notes in real time?	Integration with EMIS/SystmOne	Provides hardware?	Legal Juristiction of Product	Gives warning about low audio quality
Heidi	Until manual delete	High	Yes	Sometimes	No - few seconds after consultation	Planned	No	Victoria, Australia	No
Tortus	Current session only	Templates	Yes	No	No - few seconds after consultation	Yes	Yes - Microphones	England and Wales	No
Kiwipen	Current session, or 7 days	Templates	Yes	Optional	No - few seconds after consultation	No	No	United Kingdom	Sometimes
Nabla	14 days	High	Yes	Optional	No - few seconds after consultation	No	No	Paris, France	No
Corti Assistant	Until manual delete	Templates	Add-on	No	Yes	No	No	Florida, USA	No
Lyrebird Health	Current session, or 7 days by default	Templates	Yes	Optional	No - few seconds after consultation	No	No	Victoria, Australia	No
ConsultNote	Until manual delete	Some	Yes	Optional	No - few seconds after consultation	No	No	Victoria, Australia	No

Appendix 3 – One CareToolkit

Governance, Digital Safety and Quality Toolkit

In BNSSG, One Care, the ICB and SCW have developed an Information Governance, Digital Safety and Quality toolkit. This toolkit standardizes the approach to ascertain whether a digital tool will be safe and useful once implemented within primary care.

(*as of September 2024, until further notice the ICB and SCW are not currently involved due to staffing limitations.)

Please can the below be completed fully and sent back to digital@onecare.org.uk with any supporting documents.

Overview:

Please provide us with a link to your entry on the Companies House web page:

Please provide a brief description of your product platform used:

In simple language, please outline what you would tell GPs the benefits are of your product:

Please provide case studies from GP practices or PCNs:

Please provide a brief statement on your company's approach to Health inequalities / Net Zero / Accessibility:

What/who do you use for remote support:

Checklist

Item	Tick	If item not ticked, why?	If item ticked, please provide further detail and evidence	Copy of documentation provided?
ICO registration	□		Registration Details:	□ Yes □ No
Not escalated to ICO within last 12 months	□			□ Yes □ No
DSPT accreditation	□		Level: Date:	□ Yes □ No
ISO certification	□			□ Yes □ No
Penetration testing	□		Date: Name of company providing penetration testing: Frequency/schedule of penetration tests:	□ Yes □ No
NHS DTAC Process	□			□ Yes □ No
Any other accreditation(s) (Cyber essentials plus, etc)	□		Whole state □ Partial state □ Type of accreditation:	□ Yes □ No
Patient data is stored in UK	□			□ Yes □ No

DPIA associated with service or product	□			□ Yes □ No
Data sharing agreement and privacy notice	□			□ Yes □ No
Clinical indemnity approach	□			□ Yes □ No
Classified as a medical device	□			□ Yes □ No
Vulnerability identification and rectification process	□			□ Yes □ No
Business continuity plan	□			□ Yes □ No
Disaster recovery plan	□			□ Yes □ No

Appendix 4 - One Care Tech review discharge summary

This document has been created to explain what happens after the One Care emerging tech review process ends. The table below explains how duties and responsibilities are handed back to General Practice and the likely next steps involved for both parties.

It is important to recognise that you (the GP) bears the responsibility for issues that may arise related to information governance, patient confidentiality, clinical and digital safety in the use of this new technology. This should be seen as an advisory process, with One Care aiding practices in their due diligence, enabling them to make the most informed decisions based on the gathered information. This is particularly relevant where any potential risks relating to implementation remain and where a product interacts with the Electronic Patient Record (EPR).

Handover element	Detail
Governance and liability	The General Practice assumes responsibility for good governance, cyber insurance (if required), digital safety, etc. All information gathered during the appraisal of this supplier is attached to the discharge email and/or available on TeamNet.
Project management	Unless expressly agreed, One Care will not be responsible for managing implementation or ongoing project management of the tech product or service.
Costs	Unless expressly agreed, One Care will not be responsible for any costs relating to the tech product or service.
Digital Partner support	SCW are the current digital partner for the ICB and must be made aware of any new technology used across the NHS digital estate. They are also responsible for the issuing of the necessary usernames and required permissions for setup.
Product interaction with electronic patient record (EPR)	The practice acknowledges its responsibility for the product's interaction with the EPR and must establish the required clinical governance measures to safeguard patient safety.
Evaluation and case studies	Once the technology has been implemented, One Care would like to work with the General Practice to assess the impact. We can help you with putting measures in place, capturing impact over time, and will usually ask you to review benefit with us 12 months after the new tech has

	gone live. Unless you specifically opt-out, One Care will name your Practice in case studies when discussing the support we have provided.
Insurance	Practices should be aware that implementation of emerging tech may affect your insurance premiums. We would encourage you to talk to your insurance provider about this.