Cureus
Part of SPRINGER NATURE

Open Access Original Article

Deployment of a Cloud-Based Speech Translation Tool in a Saudi Emergency Medical Dispatch Center: A Descriptive Implementation Study

Hameed Almarwani [1], Bader Alharbi [1]

1. Communications Department, Saudi Red Crescent Authority, Madinah, SAU

Corresponding author: Hameed Almarwani, ha20034@gmail.com

Review began 10/14/2026
Review ended 11/08/2026
Published 11/16/2026

© **Copyright** 2026
Almarwani et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119773

Abstract

Background

Emergency medical dispatch depends on spoken communication, and that dependency becomes acute during mass gatherings. The Saudi Red Crescent Authority (SRCA) 997 dispatch center in Al-Madinah receives calls from pilgrims and residents who speak languages for which no on-shift interpreter is routinely available. A cloud-based speech translation tool, SMART TRANSLATION, was introduced into the dispatch workflow to support these calls. We describe what the tool is, how it was used, and what the operational record can and cannot establish about it.

Materials and methods

We conducted a retrospective descriptive analysis of an operational log of translator-assisted 997 calls and two anonymous cross-sectional dispatcher surveys. SMART TRANSLATION (versions 11 and 12) went live on 18 May 2025. The merged log held 269 records. We analyzed the 79 records that carry a complete calendar date in 2025, which span 19 May to 14 October 2025 (first to last valid record), and excluded 190 records with no date (169), a day and month without a year (16), or a year of 2015 or 2016 (5). A retrospective survey of pre-deployment experience and a post-deployment survey were administered on 14 to 16 October 2025, after deployment, and were completed by every eligible dispatcher (15 of 15 and 10 of 10). Analysis was descriptive and scripted in Python 3.10.11.

Results

All 79 analyzed records came from Al-Madinah. Six named caller languages appear, led by Urdu (44 records, 55.7%) and English (nine records, 11.4%). A further 13 records (16.5%) carry no usable language label. A dispatcher-coded field recorded instructions as given in 63 records (79.7%) and not given in 16 (20.3%). This field is a documentation category, not a measure of instruction delivery, and it was coded inconsistently. Identical free-text descriptions were coded both ways in 20 of the 79 records, and the recorded rate moved from 69.6% to 93.9% between two date strata, in a period when staff were asked to log every case.

Out of the 16 records coded as no instruction given, 11 (68.8%) describe calls in which the chief complaint was confirmed through translation and no instruction phase was recorded, and one (6.3%) records difficulty understanding the caller. The two surveys use different response bins for the understanding-time item and are not compared. All 10 post-deployment respondents chose an improvement option on an item that offered no option for no change or for worsening, and seven recommended continued use.

Conclusions

A cloud speech translation tool was in documented operational use at one Saudi dispatch center, and the dated record describes the languages of the translator-assisted calls that dispatchers chose to document. These preliminary, setting-specific findings do not establish translation accuracy, call-handling efficiency, patient safety, or patient benefit, and we make no such claim. The contribution is a characterized deployment, an account of the governance gaps that accompanied it, and an explicit list of the measurements a prospective evaluation would require.

Categories: Emergency Medicine, Health Policy, Healthcare Technology
Keywords: emergency medical dispatch, health informatics, language barriers, machine translation, speech recognition

Introduction

Emergency medical dispatch is conducted almost entirely through speech. A dispatcher must establish what has happened, where it has happened, and whether the patient is conscious and breathing, and must often deliver pre-arrival instructions, all through a telephone call with a distressed stranger. When a dispatcher

and a caller share no language, every one of those steps is degraded. Studies of emergency communication with callers of limited local-language proficiency have documented longer call handling, greater reliance on improvised workarounds, and reduced confidence in the information obtained [1]. For the most time-critical intervention in dispatch, telephone-assisted cardiopulmonary resuscitation (CPR), language discordance has been associated with measurable delay in initiating chest compressions [2], and cardiac arrest outcomes deteriorate rapidly with elapsed time [3]. The design of the dispatch protocol itself has been associated with how often callers with limited English proficiency receive telephone-assisted CPR [4].

Saudi Arabia presents this problem at an unusual scale. The Hajj and the year-round Umrah concentrate large numbers of pilgrims from across the Muslim world into a small area over short periods. In the 2025 Hajj season (1446H), 1,673,230 pilgrims performed the Hajj, of whom 1,506,576 arrived from outside the Kingdom [5]. The Saudi Red Crescent Authority (SRCA) operates the national 997 emergency medical number, and its Al-Madinah regional center receives calls from pilgrims who speak neither Arabic nor English. Reviews of health services during the Hajj have repeatedly identified communication as a limiting factor in preparedness and response [6], but the dispatch center itself is thinly represented in that literature.

Technology responses to language barriers in emergency care have been studied mostly at the point of patient contact rather than at the dispatch console. Turner et al. compared a domain-specific fixed-phrase tool with a general-purpose translation service in scenario-based testing with emergency personnel and participants of limited English proficiency and found that the domain-specific tool was preferred and more usable, while noting that communication remained constrained with either tool [7]. A participatory development program in Germany produced a fixed-phrase paramedic application covering 18 languages [8], whose usability was subsequently assessed as borderline acceptable [9] and whose effect on communication quality was examined in a nonrandomized controlled pilot study [10,11]. Speech-enabled phraselators have been compared against fixed-phrase alternatives in emergency anamnesis [12]. Each of these operates in the field, where the clinician can see the patient. None addresses the dispatch console, where the only channel is audio.

Work on automatic speech recognition (ASR) in emergency call centers has concentrated on monolingual settings. Systems have been developed for call-center speech in Azerbaijani [13] and for Romanian emergency services [14], the latter subsequently extended with a transformer-based recognizer [15]. Downstream tasks such as incident classification [16], severity prediction [17] and call-handler decision support [18] have all been studied on single-language corpora. Where multilingual emergency communication has been engineered end-to-end, evaluation has been reported on synthetic or laboratory material [19]. Recent benchmarking work makes the difficulty concrete: transcription quality for emergency dialogue degrades sharply as acoustic conditions worsen [20], and comparative evaluation of commercial recognizers on emergency medical simulation audio found weak performance precisely on the clinically loaded categories of medication and treatment [21].

What is missing from this literature is an account of what happens when such a tool is actually placed on a dispatch console and left there. This report supplies one. Its primary objective is to describe the deployment of the SMART TRANSLATION tool at the SRCA Al-Madinah 997 center and its documented operational use between 19 May and 14 October 2025. Its two secondary, descriptive objectives are to characterize the caller languages recorded in the dated operational log and to report dispatchers' perceptions of the tool from two surveys. The study is therefore a descriptive evaluation of implementation and perceived usability. It does not assess translation accuracy, call-handling efficiency, or clinical effectiveness, because the data collected during routine operations cannot support such claims. The operational records are a selected subset of documented translator-assisted calls, not the full volume of calls handled by the center, and the language distribution reported here must be read with that in mind. We also set out the measurements that a prospective evaluation would need.

Materials And Methods

Study design and reporting

This is a retrospective descriptive study of an operational record generated during routine service, combined with two anonymous cross-sectional dispatcher surveys. It is an implementation report and follows the Standards for Reporting Implementation Studies (StaRI) statement [22], and the two web surveys are reported according to the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [23]. Completed StaRI and CHERRIES checklists are provided as supplementary files. StaRI items that concern the effectiveness of the intervention have no content in a study that measures no effect and are marked as not applicable. No experimental manipulation, randomization, blinding, or prospective outcome measurement was performed. There was no control group and no pre-registered endpoint. The study was not designed in advance. It describes a deployment that had already occurred and a record that had already been collected for operational rather than research purposes. That distinction governs every claim made below.

Setting

The SRCA operates Saudi Arabia's 997 emergency medical number. The Al-Madinah regional dispatch center

serves the city of Al-Madinah and surrounding areas and experiences pronounced seasonal load during the Hajj and Umrah, when the resident population is joined by pilgrims who predominantly speak neither Arabic nor English. The study period defined below spans the 1446 Hajj season, which fell in June 2025, and the year-round Umrah traffic on either side of it. Date coverage in the record is too sparse to attribute any part of it to the pilgrimage itself, and no such attribution is made: one valid record is dated in June 2025. The record also cannot distinguish pilgrims from resident callers. Saudi Arabia has large resident communities from South Asia, and the 2022 census counted 2,116,192 Bangladeshi and 1,814,678 Pakistani residents [24]. Much of the Urdu and Bengali demand in the record may therefore come from residents rather than pilgrims. Dispatchers are Arabic speakers, most with working English. Before the deployment described here, calls in other languages were handled by seeking a colleague who happened to share the caller's language, by attempting direct communication, or by escalating.

The deployed system

SMART TRANSLATION is a Windows desktop application written in Dart using the Flutter framework. It is installed on the dispatcher workstation and operated alongside, not integrated with, the computer-aided dispatch (CAD) system. The application was developed by the two authors of this report, who are employed by the service that deployed it. Versions 11 and 12 of the application were in use during the study period, and the tool went live on 18 May 2025. The source-code audit reported below was performed on the archived source tree, whose project file carries the version string 1.3.0. That numbering differs from the numbering of the deployed builds, so the statements below about the code describe the archived source and are not claimed to hold identically for each deployed build. Figure *1* shows the pipeline as implemented.

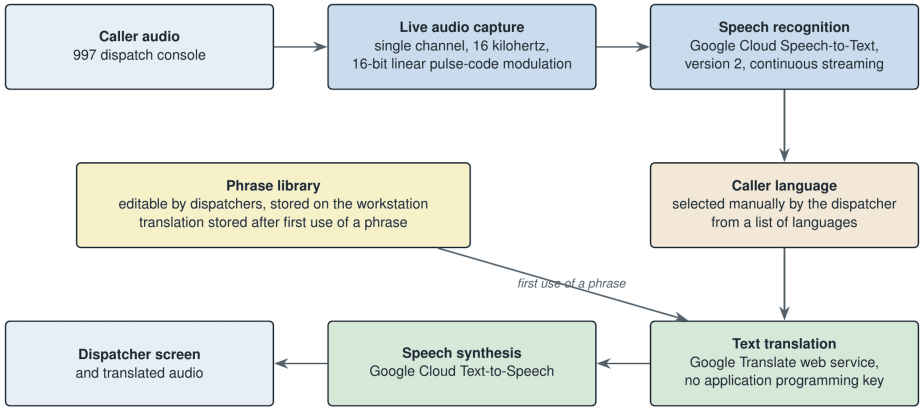


FIGURE 1: Architecture of the SMART TRANSLATION application as deployed at the SRCA 997 dispatch center

Speech recognition, text translation, and speech synthesis are performed by third-party cloud services provided by Google LLC (Mountain View, California, United States), and the caller's language is selected manually by the dispatcher. A phrase from the dispatcher-editable phrase library is translated on its first use in a language and replayed from local storage on later uses. A separate one-shot path, which submits a recorded audio file to the Speech-to-Text version 1 representational state transfer (REST) endpoint and passes the transcript to the same translation step, is not drawn. SRCA: Saudi Red Crescent Authority; 997: the national emergency medical telephone number. Figure drawn with the TikZ package (Portable Graphics Format version 3.1.10) in pdfLaTeX, TeX Live 2024.

Caller audio is captured from the console as single-channel 16 kHz LINEAR16 pulse-code-modulated audio and streamed over gRPC (a remote procedure call framework), through the google_speech Dart package (version 5.3.0), to the Google Cloud Speech-to-Text v2 service using its endless-streaming interface with the long recognition model [25]. Interim and final recognition results are both requested, and server-side voice-activity events are enabled. The session restarts every 60 seconds with a two-second transition buffer, and the client retries a failed stream up to five times before tearing the session down. Recognized text is translated through the translator Dart package (version 1.0.3+1), which calls the undocumented translate_a/single web endpoint on translate.googleapis.com without an API key. This is not the documented Cloud Translation API. The translated text is synthesized through the Google Cloud Text-to-Speech service [26], using the cloud_text_to_speech Dart package (version 1.1.3), for playback to the caller. A separate one-shot path submits a recorded audio file to the Speech-to-Text v1 REST endpoint and passes the transcript to the same translation step. All of these services are provided by Google LLC, Mountain View, California, United States, and are reached over the public internet. Speech-to-Text and Text-to-Speech were used through an SRCA institutional Google Cloud account under Google's standard Cloud terms. Recognition requests are sent to the service's "global" location, with no regional endpoint, so processing was not restricted to Saudi Arabia or to any named region, and the optional data-logging feature of the recognition service is not enabled. All connections are encrypted in transit. The translation endpoint has no published

data-processing terms and is covered by no agreement, so whether the text sent to it is retained is unknown.

The caller's language is not detected automatically. The dispatcher selects it from a list, and the application rewrites two locales that the recognition service supports only in specific countries, mapping Urdu to the Pakistan locale and Bengali to the Bangladesh locale. A phrase library of frequently used instructions is held in a local SQLite table with a title and a description field, and dispatchers create, edit, and delete its entries. The first time a phrase is played in a given language, its description passes through the same translation path as free speech and the result is stored in a second local table. Later uses of that phrase in that language replay the stored translation without calling the translation service. The application has no function for viewing or correcting a stored translation. The application stores no call audio and no transcripts. Its only local storage is the phrase library and the stored phrase translations, and audio recorded for the one-shot recognition path is held in a temporary folder that is deleted when the dispatcher starts a new call. If the cloud service fails or the connection drops during a call, dispatchers return to the methods used before deployment: seeking a colleague who shares the caller's language, attempting direct communication, or escalating the call.

Three points about the implementation matter for interpreting what follows. First, all speech recognition, translation, and synthesis are performed by third-party cloud services reached over the public internet. Second, the application contains no speech enhancement, noise suppression, confidence thresholding, terminology glossary, automated language identification, or automated removal of identifying information. Each of these is absent from the source code. Third, the application performs no timing instrumentation, so no measurement of end-to-end latency exists for the deployed system.

Study period, sampling approach, and eligibility of records

Sampling Approach

No sampling was performed on the case log. Every record in the merged operational log supplied to the investigators (269 records) was screened against the criteria below, and every record that met them was analyzed. The analyzed set is therefore the complete set of eligible records and not a probability sample of translator-assisted calls. The log itself contains the cases that dispatchers chose to record. It is not a register of every call handled, and total call volume is unknown.

Inclusion Criteria

A record was included if it (1) documents a translator-assisted call to the SRCA 997 emergency medical number and (2) carries a complete calendar date, with day, month, and year, in which the year is 2025, the year of deployment.

Exclusion Criteria

A record was excluded if its date field was empty (169 records), held a day and month with no year (16 records), or held a year of 2015 or 2016 (five records). Years of 2015 or 2016 cannot be correct for a deployment that took place in 2025. The weekday recorded with each of these five records also does not match its stated date (1 June 2015 was a Monday and 1 June 2016 a Wednesday). The three records marked "Sunday" match 1 June 2025, which was a Sunday, so these records are most likely year-entry errors made during the deployment rather than records that predate the tool. The five records were excluded and were not corrected, because assigning a corrected year would require an assumption that the source does not support. All 169 records transcribed from a WhatsApp group carry no date field and are therefore all excluded, so every analyzed record comes from the structured spreadsheet.

Study Period

The study period is the interval from the date of the first valid record to the date of the last valid record: 19 May 2025 to 14 October 2025, which is 149 calendar days and 31 distinct dates. This interval was derived from the data after the exclusions had been applied. It was not fixed in advance and it is not the period of deployment. SMART TRANSLATION went live on 18 May 2025, one day before the first valid record. Statements in this report about the study period refer to this interval.

Data sources

Three files constitute the entire data basis for this report.

The merged case log contains 269 records of translator-assisted 997 calls and merges two streams with different structures. One hundred records originate from a structured spreadsheet maintained by dispatchers, which includes a field recording whether instructions were given. The spreadsheet was created at go-live in May 2025 to log calls handled with SMART TRANSLATION, and the "interpreter" named in its action fields is the dispatcher operating the tool. It has no separate field confirming tool use for each call,

and no call audio or application log is available to verify it record by record. One hundred and sixty-nine records were transcribed from a WhatsApp group used by shift staff to note translator-assisted cases and carry no date and no instruction-outcome field. The fields of the structured stream are a case number, region, caller language as recorded, one free-text case note, two categorical interpreter actions, an instruction-outcome value, a day of the week, and a date. The day of the week is filled in for 20 of the 79 analyzed records, and in all 20 it agrees with the recorded date. No free-text value in the analyzed set is truncated.

Language of the Record and of This Analysis

The operational record is kept in Arabic. The analysis was performed on an English version of each of the three workbooks. That version was used as supplied to the investigators. No back-translation or other verification step was applied to it, so we make no claim that the English rendering was independently verified. The English merged log documents its own editing in a summary sheet: language spellings were normalized so that variant spellings of one language map to a single label, region spellings were normalized in the same way, and dates written as full dates were converted to date values while 16 records carrying only a day and month were left as text with no year assumed. A column holding the value originally typed by the agent is retained and was inspected during this analysis. It contains English glosses and editorial annotations rather than the original Arabic string. The summary sheet states that eight records had a medical complaint entered in the language field. The cell-level count is 12, and all 12 are among the analyzed records.

Two dispatcher surveys were administered anonymously through Google Forms (Google LLC, Mountain View, California, United States). Both were run as an internal service evaluation of the tool, before any decision to publish. The retrospective survey of pre-deployment experience (seven items) collected responses between 14 and 15 October 2025, after the tool had been in routine use since May 2025, and asked dispatchers to recall communication with non-Arabic-speaking callers before the tool was available. The post-deployment survey (nine items) collected responses between 15 and 16 October 2025 and asked about experience using the tool. The two administration windows overlap: both forms were open simultaneously for roughly seven hours on 15 October. Both instruments were administered in Arabic. The retrospective survey form presented one of its items, the effect of language on the speed of medical instructions, twice. Fourteen of the 15 respondents gave the same answer both times. One answered "sometimes" and then "yes." The first answer is the one analyzed here, which gives 10 "yes" and five "sometimes"; the repeated item would give 11 and four.

The analysis used English-language versions of the two response exports. The original Arabic exports were checked against them: respondent counts (15 and 10), timestamps, and every answer agree, and within each item each Arabic option corresponds to exactly one English option. Screenshots of the Google Forms response summaries show the same respondent counts and the same distributions for the first two items of each form. Appendices A and B reproduce the items in Arabic with English translations and list every response option offered, taken from the original forms. Two post-deployment items offered only positive options. The accuracy item offered no option below "moderately accurate," and the improvement item offered only "significant improvement" and "slight improvement." The improvement item is therefore leading, and both items are reported with that limitation.

Development of the Instruments

One author (HA) wrote the items of both instruments. The project documents no pilot test, expert review, or validation of either instrument. That author also developed SMART TRANSLATION, so the post-deployment items were written by a developer of the tool they ask about. The same author distributed the forms to colleagues in the same service, which may have made some feel obliged to take part or to answer favorably.

Eligibility, Sampling, and Response Rate

No sample was drawn: each survey was open to all of its eligible dispatchers. The retrospective survey of pre-deployment experience was open to all dispatch staff, to describe the situation before the program was used. The post-deployment survey was open only to staff who had used the program. This difference in definition explains the difference in group size: 15 eligible dispatchers for the retrospective survey and 10 for the post-deployment survey. The five dispatchers who had not used the tool preferred the methods they used before deployment, asking a colleague for help or communicating with the caller directly. This reason comes from the authors' knowledge of the center and was not collected systematically. Every eligible dispatcher responded, giving response rates of 100% (15 of 15) for the retrospective survey and 100% (10 of 10) for the post-deployment survey. Because every eligible dispatcher took part, each survey is a census of its eligible group and not a random sample. The instruments were anonymous, so the response rates rest on the authors' count of eligible dispatchers and cannot be verified against individual invitations. Participation was voluntary and uncompensated, and responses were not linked to individuals or to each other.

Analysis

All counts were produced by a scripted pipeline that reads the three source workbooks and emits a machine-readable results file, from which every table and figure in this manuscript is generated. The pipeline was written in Python version 3.10.11 (Python Software Foundation, Wilmington, Delaware, United States) using the openpyxl library version 3.1.5 to read the workbooks. The bar chart was drawn with Matplotlib version 3.8.2 [27], and the diagrams with TikZ. No value was transcribed by hand. The pipeline, and a verification script that re-checks the numbers appearing in the manuscript text against that results file, are available from the corresponding author on request.

Three conventions are applied throughout and are stated here because they determine the denominators and the wording of every result.

First, all 79 analyzed records carry the instruction-outcome field, so every outcome statistic is computed over the 79 analyzed records. The 169 records without the field are WhatsApp-derived and were excluded for lack of a date. They are not counted as failures, because that would manufacture a failure rate out of a structural difference between two record streams.

Second, that field is treated throughout as an operational coding decision and not as a service-performance measurement. It records what a dispatcher categorized the episode as during routine work. No written coding rule for it exists in any project document, no call audio, timestamp, or independent observation underlies it, and the Results report three checks on its reliability. Wording in this report is chosen accordingly: the record states that instructions were recorded as given, not that delivery was measured.

Third, the two surveys are reported as two anonymous cross-sectional surveys of dispatchers at one center. Responses cannot be linked between the instruments, statistical independence is not assumed, and the distributions are never differenced, compared by calculating ratios, or tested against one another. The reason is specific: the understanding-time item uses different response bins in the two instruments. The fastest option before deployment is "less than two minutes" and the fastest option after deployment is "less than one minute." A dispatcher whose performance did not change at all would select different bins on the two forms, so any apparent shift is a property of the instrument rather than of the system. Both instruments were also administered in October 2025, after the tool had entered routine use, which is why the first is called a retrospective survey of pre-deployment experience: it measures recall, not a baseline. We report both response distributions in full and draw no comparative inference from them. The same recall limitation applies to the post-deployment item that asks dispatchers to rate the improvement compared with before the tool. That item is a retrospective self-comparison answered by dispatchers of the service that deployed the tool, and it is reported as a distribution of judgments, not as evidence of change.

Free-text case notes for records coded as no instruction given were classified by matching against an ordered set of patterns corresponding to the outcomes the notes describe, with the more clinically salient pattern taking precedence where a note describes more than one. The pattern set and its precedence order are part of the analysis pipeline, which is available on request. Because the note text is short, formulaic, and frequently repeated verbatim, this classification summarizes what was written rather than adjudicating what occurred.

Descriptive statistics are counts and percentages, with percentages rounded to one decimal place (survey percentages to whole numbers), so that totals may differ from 100 by rounding. No inferential test is reported anywhere in this study. The approach follows from the study objectives, which are descriptive. Three further features of the data point the same way: the log contains the cases dispatchers chose to record and is not a probability sample, the instruction-outcome field is an operational coding decision rather than a measurement, and no comparative question was specified in advance. The small numbers involved (79 outcome-bearing records across six named languages and one unlabeled stratum, and survey groups of 15 and 10) are an additional consideration. The post hoc split of records into two date strata reported in the Results was chosen after the monthly counts had been inspected and is not a prespecified contrast.

Ethical considerations

The analysis uses operational service records and anonymous voluntary staff surveys. The case log contains no patient names, telephone numbers, addresses, or clinical identifiers. Records are identified by an internal case number and describe the interpreter's action rather than the patient's condition in detail. Survey responses were not linked to individual dispatchers.

The study was reviewed by the SRCA Institutional Review Board (IRB), which is registered with the Saudi National Committee of Bioethics under number H-01-R-110. The application was submitted on 3 November 2025 under the study title "Cloud and AI-Powered Real-Time Translation Program for Saudi Emergency Call Centers", and on 30 December 2025 the board approved the study in the exempt category (IRB log number IRB-25-0042S). The application followed the collection of both data sources and stated that they had already been collected. The analyzed case records are dated 19 May to 14 October 2025, and the surveys were administered on 14 to 16 October 2025 as an internal service evaluation of the tool, before any decision to publish. The board's determination was therefore made on data that it knew had already been collected.

The surveys were administered anonymously through an online form during routine operations. The project holds no written participant information sheet issued to dispatchers at the time, and no signed consent forms. What is documented is that participation was voluntary and uncompensated, that no identifying information was collected, and that a submitted response was the only act of participation. Consent was therefore inferred from voluntary completion rather than recorded.

One governance finding emerged from the data itself and is reported here because it bears on the conduct of the service rather than on the analysis. One hundred and sixty-nine of the 269 records in the merged log, 62.8% of it, were transcribed from a WhatsApp group used by shift staff. Operational detail about emergency cases therefore passed through a consumer messaging platform that is not an approved clinical record system. No patient identifiers appear in the transcribed records available to us, but the practice is not consistent with appropriate handling of operational health information and warrants review independently of this study. It is discussed with the governance findings that concern the tool itself under Data protection and clinical governance in the Discussion.

Results

Selection and completeness of the operational record

The merged log holds 269 records. Table *1* and Figure *2* show how they were reduced to the analyzed set. One hundred and ninety records (70.6%) were excluded: 169 carry no date at all, 16 give a day and month with no year, and five carry a year of 2015 or 2016. The 169 records without a date are exactly the 169 WhatsApp-derived records, so the exclusion removes the whole WhatsApp stream and 21 of the 100 structured records. The 79 analyzed records (29.4%) all come from the structured spreadsheet, all carry the instruction-outcome field, and all originate from the Al-Madinah region.

Record group	Role in analysis	Records	Percentage of 269 (%)
Records in the merged operational log	Starting set	269	100
Records from the structured spreadsheet log	Starting set	100	37.2
Records from the WhatsApp-derived log	Starting set	169	62.8
No date recorded	Excluded	169	62.8
Day and month recorded, no year	Excluded	16	5.9
Year recorded as 2015 or 2016	Excluded	5	1.9
Total excluded	Excluded	190	70.6
Valid calendar date in 2025	Analyzed	79	29.4

TABLE 1: Derivation of the analyzed set from the merged 997 translator-assisted case log

Percentages are of all 269 records in the log and are rounded to one decimal place. The WhatsApp-derived log carries no date field, so all of its records are excluded.

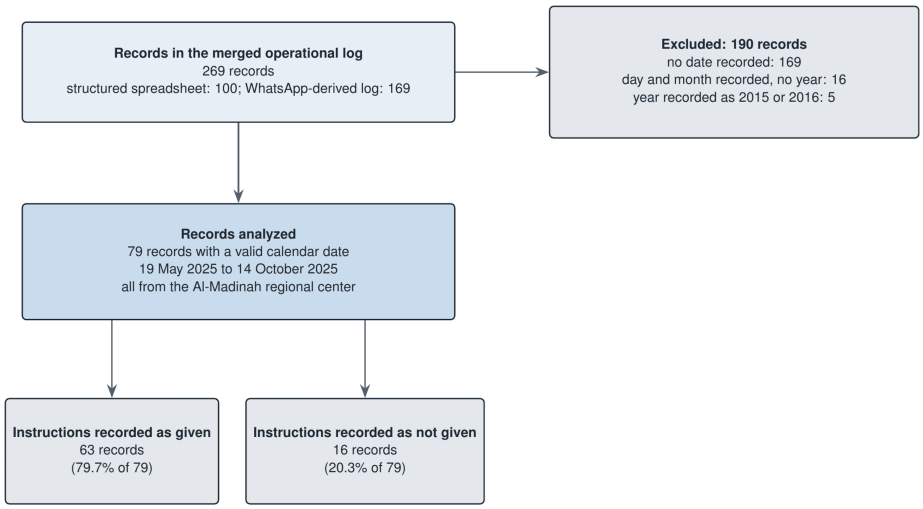


FIGURE 2: Derivation of the analyzed set from the merged operational log of translator-assisted calls

Calls were made to the Saudi Red Crescent Authority 997 emergency medical number. Records with no date, with a day and month but no year, or with a year of 2015 or 2016 were excluded. Percentages are of the 79 analyzed records. Figure drawn with the TikZ package (Portable Graphics Format version 3.1.10) in pdfLaTeX, TeX Live 2024.

The first valid record is dated 19 May 2025 and the last 14 October 2025, so the study period spans 149 calendar days, of which 31 dates carry at least one record. Records are distributed as May 12, June one, July two, August five, September 26, and October 33. The record therefore cannot support a monthly time series, a trend estimate, or any statement about how call volume changed over the period, and we report none. Nor can it be localized to the pilgrimage itself. The 1446 Hajj season fell in June 2025, and exactly one valid record is dated in that month. The tool's use during the Hajj, as distinct from year-round Umrah traffic, therefore cannot be measured from these data.

The exclusion was not neutral. Table *2* shows that the excluded records differ from the analyzed ones: English accounts for 26.8% of the excluded records and 11.4% of the analyzed records, eight language labels, covering seven distinct languages (Indonesian, Hindi, Uzbek, Malay, Persian, Swahili, Urdu/Hindi, and Hindi/Malayalam), occur only among the excluded records, and all 17 records from other regions and both records with no region recorded were excluded. Findings about the analyzed set should not be read as findings about the whole log.

Value in the caller-language field	Category	Analyzed records	Analyzed (% of 79)	Excluded records	Excluded (% of 190)
Urdu	Named language	44	55.7	104	54.7
English	Named language	9	11.4	51	26.8
Bengali	Named language	6	7.6	10	5.3
Indonesian	Named language	0	0	8	4.2
Turkish	Named language	3	3.8	4	2.1
Hindi	Named language	0	0	4	2.1
Russian	Named language	3	3.8	0	0
Uzbek	Named language	0	0	2	1.1
French	Named language	1	1.3	0	0
Hindi/Malayalam	Named language	0	0	1	0.5
Malay	Named language	0	0	1	0.5
Persian	Named language	0	0	1	0.5
Swahili	Named language	0	0	1	0.5
Urdu/Hindi	Named language	0	0	1	0.5
Subtotal, named languages	Subtotal	66	83.5	188	98.9
Medical complaint or generic case descriptor entered	No usable label	12	15.2	0	0
Field left blank	No usable label	1	1.3	2	1.1
Subtotal, no usable label	Subtotal	13	16.5	2	1.1
Total	Total	79	100	190	100

TABLE 2: Values recorded in the caller-language field of the merged case log, by analyzed and excluded records

Named languages are separated from records with no usable language label. Percentages are of the number of records in each group (79 analyzed, 190 excluded). A medical complaint or a generic case descriptor was typed into the language field instead of a language in the 12 records so marked.

Caller languages

Six named caller languages appear among the 79 analyzed records (Table *2*, Figure *3*). Urdu accounts for 44 records (55.7%) and English for nine (11.4%). Bengali follows with six records (7.6%), then Turkish and Russian with three each (3.8% each) and French with one (1.3%).

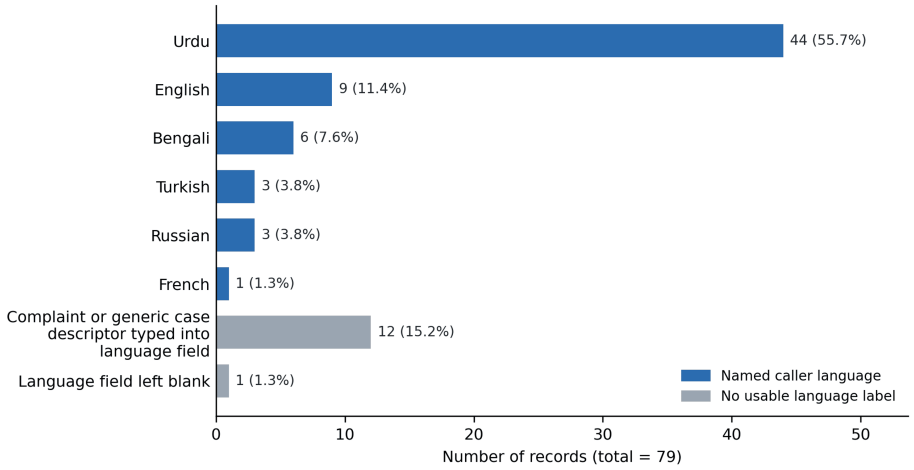


FIGURE 3: Distribution of values recorded in the caller-language field across the 79 analyzed records

Percentages are of the 79 records. Blue bars are named caller languages and grey bars are records with no usable language label. Figure drawn with Matplotlib version 3.8.2 under Python version 3.10.11.

A further 13 records (16.5%) carry no usable language label, and they are reported as a separate category rather than as a seventh language. In 12 of them a medical complaint or a generic case descriptor was typed into the language field: four read "medical case" and one read "sick person", which identify no complaint at all, while the remaining seven are specific, comprising breathing problems (two), cardiac and respiratory arrest (two), chest pain, headache and scorpion sting. In the other record the field was left blank.

These 13 records are 16.5% of the analyzed set, where they carry more weight than that percentage suggests. The point is taken up in the outcome analysis below. The language distribution should also not be read as the linguistic profile of 997 demand in Al-Madinah. It is the profile of calls at which a translator-assisted workflow was invoked, a record was made, and a valid date was entered, which is a narrower thing.

Geographic distribution

All 79 analyzed records originate from the Al-Madinah regional center. Records from four other regions (Tabuk, nine; Al-Jouf, four; Northern Borders, three; and Eastern Province, one) occur only in the excluded WhatsApp-derived stream, so the analyzed set describes a single center.

Interpreter actions

The dominant recorded activity is establishing what is wrong. The first interpreter action is recorded as the chief complaint in 68 of the 79 records (86.1%), as "all of them" in six (7.6%), as giving instructions in two (2.5%), as CPR in two (2.5%), and as key questions in one (1.3%). This field is a categorical choice and no value is truncated, so these counts are complete.

A second interpreter action is recorded for all 79 records. It was giving instructions in 50 (63.3%), a residual "other" category in 21 (26.6%), locating the incident in two, adjusting the case coding in two, and key questions, CPR, chief complaint, and "all of them" in one each.

CPR appears in three records across all free-text and categorical fields, and all three carry the outcome field. In one Urdu record, CPR instructions were begun and the caller reported the ambulance arriving before compressions started. One Bengali record states that instructions were given until the ambulance crew arrived. The third, also Bengali, states that the caller refused the instructions although he received them correctly through translation. It is coded as instructions given, and it records a caller declining resuscitation guidance after receiving it rather than guidance being carried out. Two further records, dated 12 and 14 October 2025, have "cardiac and respiratory arrest" typed into the language field. Their notes read "Confirming the case and giving the instructions" and "Confirming the chief complaint," and the second interpreter action in both is giving instructions, but neither names CPR or records the caller's language. The record therefore cannot show which instructions were given in these two calls or in what language. These five records cannot support any statement about the tool's performance during cardiac arrest, and we make none. No incidents or near-misses involving the tool were reported through the service's safety-reporting system during the study period.

Instructions recorded as given: a documentation category

In 63 of the 79 analyzed records (79.7%), a dispatcher recorded that instructions were given, and in 16 (20.3%) the dispatcher recorded that they were not. This figure is a documentation category and not an instruction delivery rate. The records cannot separate instructions that were indicated, attempted, communicated, understood, and followed, because none of these stages are recorded for most calls. The field is a dispatcher's own categorization of the episode, entered during routine service with no written coding rule anywhere in the service's documentation. No call audio, system timestamp, or independent observation underlies it. It is an operationally recorded outcome and not a measurement of the tool's performance, and three checks reported below show how far apart those two things are here.

The figure is not homogeneous across the study period (Table *3*). The recorded rate is 69.6% among the 46 records dated 19 May to 30 September 2025 and 93.9% among the 33 dated 1 to 14 October 2025. This split was made after the monthly counts had been inspected and is not a prespecified contrast. Even so, a 24-point spread across two adjacent date strata of comparable size is not what a stable service measurement looks like, and the pooled value should not be read as one. Recording practice changed in the second stratum. In the weeks before the service evaluation, staff were asked to log every translator-assisted case, and some records were entered retrospectively rather than at the time of the call. The records show the same change: all 12 records with a medical complaint typed into the language field are dated 1 to 14 October 2025, and all 12 are coded as given. Among records with a named language, the recorded rate is 71.1% (32 of 45) from 19 May to 30 September and 90.5% (19 of 21) from 1 to 14 October. The higher October rate therefore reflects, at least in part, a change in recording practice and should not be read as a change in the service.

Period of the record date	Records	Recorded as given	Recorded as not given	Recorded as given (%)
19 May to 30 September 2025	46	32	14	69.6
1 October to 14 October 2025	33	31	2	93.9
All analyzed records	79	63	16	79.7

TABLE 3: Instruction outcome recorded by the dispatcher, by date stratum

Data are for the 79 analyzed records, split into two date strata. The split is post hoc. Percentages are of the records in each row.

The second interpreter action corroborates the field only in part. That field holds a single value for each record, drawn from eight categories that include "all of them," so a dispatcher recorded one action per call. "Giving instructions" is the second action in 49 of the 63 records coded as given. The other 14 record a different action: the residual "other" category in eight, locating the incident in two, and key questions, CPR, chief complaint, and adjusting the case coding in one each. Because the field records one action only, these 14 leave the "given" code unconfirmed rather than contradicting it. A direct contradiction requires a record coded as not given whose second action entails that instructions were given. That occurs in two of the 16 records coded as not given: one whose second action is "giving instructions" and one whose second action is "all of them."

Identical case descriptions were coded both ways in 20 records, a finding set out in full at the end of this section.

Table *4* breaks the field down by caller language, and two features of that breakdown limit what can be read from it. First, only Urdu (44 records) reaches double figures. English has nine records, Bengali six, Turkish and Russian three each, and French one. Rates computed on nine or fewer records, and especially on three or one, are reported for completeness and support no per-language conclusion. Second, eight language labels in the full log, covering seven distinct languages (Hindi, Indonesian, Malay, Persian, Swahili, Uzbek, Urdu/Hindi, and Hindi/Malayalam), have no analyzed record and therefore no outcome statistic at all.

Value in the caller-language field	Category	Records	Recorded as given	Recorded as not given	Recorded as given (%)
Urdu	Named language	44	32	12	72.7
English	Named language	9	8	1	88.9
Bengali	Named language	6	5	1	83.3
Turkish	Named language	3	2	1	66.7
Russian	Named language	3	3	0	100
French	Named language	1	1	0	100
Medical complaint or generic case descriptor entered	No usable label	12	12	0	100
Field left blank	No usable label	1	0	1	0
Total	Total	79	63	16	79.7

TABLE 4: Instruction outcome recorded by the dispatcher, by caller language

Data are for the 79 analyzed records. Percentages are of the records in each row. A medical complaint or a generic case descriptor was typed into the language field instead of a language in the 12 records so marked, and one record has a blank language field.

A third feature belongs with them. The 13 records with no usable language label are not a language row and are reported separately in the table, but they are 16.5% of the analyzed set and supply 12 of the 63 records coded as given, which is 19.0% of that numerator. All 12 records with a complaint or a generic descriptor in the language field are coded as given, while the one record with a blank field is coded as not given. About a fifth of the records coded as given therefore come from records in which nobody recorded what language the caller spoke.

Within those constraints, the recorded rate for Urdu is 72.7% (32 of 44). The rates for the five smaller languages range from 66.7% to 100.0% on between one and nine records each, and a single additional record coded the other way would move any of them by 11 points or more, so no ordering by language is drawn from these data.

Free-text notes of the records coded as no instruction given

The 16 records coded as no instruction given were examined individually against their free-text notes (Table *5*).

What the free-text note records	Records	Percentage of 16 (%)
Chief complaint confirmed through translation, no instruction phase recorded	11	68.8
Caller declined the instructions after receiving them correctly	1	6.3
Low-acuity disposition, caller referred to another service	1	6.3
Ambulance crew arrived before the interaction was completed	1	6.3
Caller did not respond	1	6.3
Dispatcher recorded difficulty understanding the caller	1	6.3
Total	16	100

TABLE 5: Classification of the 16 records coded as no instruction given

Records were classified against their free-text dispatcher notes. Percentages are of the 16 records and, because of rounding, sum to slightly more than 100.

In 11 of the 16 (68.8%), the note records that the chief complaint was confirmed or identified through translation, with no instruction phase described. In 10 of those 11, the second action is coded "other," and in one it is "all of them." That last value includes giving instructions, so it contradicts the outcome code on the same row. It is one of the two direct contradictions reported above, and it is the reason the second action is described here by its distribution rather than summarized with a single value. These 11 notes do not describe the tool failing to convey instructions. They describe calls in which the recorded interaction ended with triage. The remaining five each occur once. One caller declined the instructions after receiving them correctly through translation. One case received a low-acuity disposition, and the caller was referred to another service. In one, the ambulance crew arrived before the interaction was completed. In another, the caller did not respond. The last states that there was difficulty understanding the caller, in a case that was subsequently triaged as low acuity.

That last record is the only one in the analyzed set that attributes a non-delivery to a comprehension problem. This is a documented negative result rather than an absence of looking: all three free-text fields of all 79 records were searched for 19 terms covering difficulty, misunderstanding, mistranslation, error, failure, noise, dialect, accent, repetition and clarification, and that search returns exactly one record. No record attributes a non-delivery to dialect, to the resource level of a language, or to background noise. Although language discordance and dialect variation are frequently hypothesized as causes of dispatch failure, the operational notes contain no documentation of such factors, and no speculative root-cause attribution is made.

The third check completes the picture. Two distinct note texts appear under both outcome codes, together covering 20 of the 79 analyzed records. The note "the chief complaint was confirmed through translation and consciousness and breathing checked" was coded as given seven times and as not given eight times. The note "the caller refused the instructions, although he received them correctly through translation" was coded as given four times and as not given once. Identical descriptions of a call therefore received opposite outcome codes. Taken with the spread across date strata and the two direct contradictions of the second interpreter action, the field records an individual dispatcher's judgment of whether the episode counted as an instruction event rather than a reproducible measurement, and the 79.7% figure should be read as a description of that coding practice at least as much as of the service.

Dispatcher surveys

The two instruments are reported separately (Tables *6*, *7*) and are not compared. Each was completed by every eligible dispatcher (15 of 15 and 10 of 10). Their administration windows overlap; responses could not be linked between them, and statistical independence is not assumed. The one item with the same response labels in both, the frequency of receiving calls in a language other than Arabic or English, gives a similar picture in each: often four, sometimes eight, rarely three before deployment, and often two, sometimes six, rarely two after. This is reported descriptively to indicate that the two groups had broadly comparable exposure, not as a test.

Survey item	Response	Respondents	Percentage of 15 (%)
Q1. Frequency of calls in a language other than Arabic/English	Often	4	27
Q1. Frequency of calls in a language other than Arabic/English	Sometimes	8	53
Q1. Frequency of calls in a language other than Arabic/English	Rarely	3	20
Q2. How the call is handled	I try to communicate directly	12	80
Q2. How the call is handled	I ask a colleague for help	3	20
Q3. Clarity of communication	Very clear	1	7
Q3. Clarity of communication	Clear	2	13
Q3. Clarity of communication	Unclear	9	60
Q3. Clarity of communication	Very difficult	3	20
Q4. Time taken to understand the case	Less than 2 minutes	6	40
Q4. Time taken to understand the case	2 to 5 minutes	6	40
Q4. Time taken to understand the case	More than 5 minutes	3	20
Q5. Effect of language on the speed of medical instructions	Yes	10	67
Q5. Effect of language on the speed of medical instructions	Sometimes	5	33
Q6. Frequency of CPR instruction attempts with non-Arabic speakers	Several times	5	33
Q6. Frequency of CPR instruction attempts with non-Arabic speakers	A few times	6	40
Q6. Frequency of CPR instruction attempts with non-Arabic speakers	Once	1	7
Q6. Frequency of CPR instruction attempts with non-Arabic speakers	Never	3	20
Q7. Callers' understanding of CPR instructions	Yes	2	13
Q7. Callers' understanding of CPR instructions	Partially	5	33
Q7. Callers' understanding of CPR instructions	No	6	40
Q7. Callers' understanding of CPR instructions	Not sure	2	13

TABLE 6: Retrospective survey of pre-deployment experience: all items and response distributions

Fifteen respondents. The survey was administered in October 2025, after deployment, and asked dispatchers to recall the period before the tool was available. Percentages are of the 15 respondents and are rounded to whole numbers. CPR: cardiopulmonary resuscitation.

Survey item	Response	Respondents	Percentage of 10 (%)
Q1. Frequency of calls in a language other than Arabic/English	Often	2	20
Q1. Frequency of calls in a language other than Arabic/English	Sometimes	6	60
Q1. Frequency of calls in a language other than Arabic/English	Rarely	2	20
Q2. Ease of use of the system	Very easy	6	60
Q2. Ease of use of the system	Easy	2	20
Q2. Ease of use of the system	Slightly difficult	2	20
Q3.Perceived translation accuracy of the Arabic output	Very accurate	4	40
Q3.Perceived translation accuracy of the Arabic output	Accurate	1	10
Q3. Perceived translation accuracy of the Arabic output	Moderately accurate	5	50
Q4. Clarity of the translated instructions to the caller	Very clear	5	50
Q4. Clarity of the translated instructions to the caller	Clear	4	40
Q4. Clarity of the translated instructions to the caller	Confusing	1	10
Q5. Time taken to understand the case	Less than 1 minute	2	20
Q5. Time taken to understand the case	1 to 2 minutes	7	70
Q5. Time taken to understand the case	More than 2 minutes	1	10
Q6. Frequency of delivering CPR instructions using the system	Several times	4	40
Q6. Frequency of delivering CPR instructions using the system	A few times	2	20
Q6. Frequency of delivering CPR instructions using the system	Once	1	10
Q6. Frequency of delivering CPR instructions using the system	Never	3	30
Q7. Clarity of CPR instructions after translation	Very clear	4	40
Q7. Clarity of CPR instructions after translation	Clear	4	40
Q7. Clarity of CPR instructions after translation	Unclear	2	20
Q8. Level of improvement compared with before the system	Significant improvement	6	60
Q8. Level of improvement compared with before the system	Slight improvement	4	40
Q9. Recommendation to use the system in 997 operations	Yes	7	70
Q9. Recommendation to use the system in 997 operations	Maybe	2	20
Q9. Recommendation to use the system in 997 operations	Not sure	1	10

TABLE 7: Post-deployment dispatcher survey: all items and response distributions

Ten respondents. Percentages are of the 10 respondents. Three of the 10 respondents had never delivered CPR instructions using the system (Q6), so only seven had direct experience of the CPR items. Among those seven, Q7 responses were very clear (four), clear (two), and unclear (one). Q3 offered only the three options shown, and Q8 offered only improvement options. CPR: cardiopulmonary resuscitation. 997: the national emergency medical telephone number.

In the retrospective survey of pre-deployment experience (15 respondents, Table 6), 12 of 15 respondents described communication with non-Arabic-speaking callers as unclear or very difficult, and 12 of 15 said they attempted to communicate directly rather than seeking help from a colleague. All 15 reported that language affected the speed of delivering medical instructions, answering "yes" (10) or "sometimes" (five). Eleven of 15 reported that callers understood CPR instructions only partially or not at all.

In the post-deployment survey (10 respondents, Table 7), eight of 10 rated the tool easy or very easy to use. On perceived translation accuracy of the Arabic output, four chose "very accurate," one "accurate," and five "moderately accurate." The item offered only these three options, so "moderately accurate" was the lowest

rating available and the item could not record a negative rating. Nine of 10 rated the clarity of translated instructions to the caller as clear or very clear. The improvement item offered only two options, "significant improvement" and "slight improvement," with no option for no change or for worsening. Six respondents chose the first and four the second. Because the item is leading and asks for a retrospective comparison, this distribution does not show that every respondent perceived improvement. It records how dispatchers answered, not a measured change. Seven of 10 recommended continued use, two answered "maybe" and one was unsure.

The two CPR items in these instruments were answered by every respondent, including those who reported no relevant experience, and the distributions are therefore reported in both forms. After deployment, eight of 10 rated the translated CPR instructions clear or very clear and two rated them unclear. But three of the 10 had answered "never" to how often they had delivered CPR instructions using the tool, and those three rated it clear, clear, and unclear. Among the seven respondents who had used it for CPR, four rated the instructions very clear, two clear, and one unclear. Before deployment, 11 of 15 reported that callers understood CPR instructions only partially or not at all. Three of the 15 had never attempted CPR instructions with a non-Arabic speaker, and among the 12 who had, nine reported partial understanding or none. We report both the full and the exposed distributions and draw no comparison between them.

These ratings come from Arabic-speaking dispatchers judging Arabic output. They record whether the Arabic the dispatcher received was usable. They are not, and cannot be, an assessment of whether the Urdu, Bengali, or Turkish source speech was rendered faithfully, since no respondent speaks those languages. The accuracy item is accordingly reported as perceived translation accuracy of the Arabic output, and no percentage accuracy scale is derived from it.

Discussion

What this deployment shows

The defensible findings of this study concern documented operational use rather than effect. A cloud speech translation tool was installed on a live dispatch console in a Saudi emergency medical center and was used on at least 31 distinct dates between 19 May and 14 October 2025, generating 79 dated records of calls in which dispatchers invoked it. The merged log holds 190 further records that could not be dated and were excluded. The languages documented in the analyzed records are concentrated and varied: six named languages, with Urdu and English together accounting for 67.1% of records and Bengali, Turkish, Russian, and French appearing in small numbers, alongside 13 records in which no usable language label was entered at all. These figures describe the translator-assisted calls that dispatchers chose to document, not the overall linguistic demand of the center. The full log is more varied still, holding 14 distinct language labels covering 13 distinct languages, but the languages beyond these six occur only among the excluded records and cannot be characterized further. The dominant recorded use was establishing the chief complaint, which is the first interpreter action in 86.1% of analyzed records.

SMART TRANSLATION runs beside the CAD system rather than within it, requires the dispatcher to select the caller's language manually, and depends on third-party cloud services over the public internet. It was in documented operational use on at least 31 distinct dates, and eight of 10 post-deployment respondents rated it easy or very easy to use. Five of the 15 dispatchers had not used it and preferred their previous methods, so adoption was not universal. Training burden, cost, failed use, downtime, and workflow disruption were not measured, and the study does not establish feasibility in any wider sense. What it does show is that the measurements needed to evaluate such a tool were not in place when it was introduced.

What this deployment does not show

This study reports no effect on call-handling time, no translation accuracy metric, and no patient outcome. Each omission is deliberate, and each has a specific cause.

Call-handling time cannot be estimated because no timing data exists. The application performs no timing instrumentation, and the dispatch system's own timestamps could not be used. CAD records are held by a separate SRCA unit; the authors have no access to CAD exports, and extracting them would add a new data source that requires a separate data-access approval and an amendment to the IRB approval. Linkage through the case number is planned as the first step of the prospective evaluation. The two surveys asked dispatchers to recall how long they took to understand a case, but they used different response bins, and both were administered after the tool had been in daily use. A dispatcher whose performance was unchanged would select different bins on the two forms purely because the fastest available option moved from "less than two minutes" to "less than one minute." Differencing those distributions would measure the questionnaire, not the system.

Translation accuracy cannot be estimated because no bilingual assessment was performed. The only accuracy-adjacent data is a single survey item answered by Arabic-speaking dispatchers about Arabic output. Whether an Urdu utterance was faithfully rendered is not a question an Arabic monolingual can answer, and the distinction matters clinically: the failure mode of concern in dispatch is a translation that is fluent and

confident and wrong. Comparative evaluation of commercial recognizers on emergency medical simulation audio found the weakest performance precisely on medication and treatment content [21], and benchmarking under realistic acoustic conditions shows transcription quality falling as noise rises [20]. Neither risk is observable in our record.

Patient outcomes were never collected. No linkage to clinical records, response intervals, or survival exists, and none was attempted. Any inference from a dispatch-level communication tool to survival would require several unvalidated steps, and we draw none.

Finally, the instruction-outcome field is less reliable than its apparent precision suggests, for the reasons set out in the Results and the Limitations. The 79.7% figure is a documentation category, not an objective measure of service performance, and nothing in this record can be substituted for such a measure.

Relation to prior work

Most evaluated language tools in emergency care operate at the point of patient contact. The German paramedic application developed through a participatory process [8] was assessed for usability [9] and then in a nonrandomized controlled pilot study that reported improved perceived communication quality alongside longer time on scene [10]. A subsequent correction revised that study's numerical results [11], and we therefore cite its direction rather than its magnitudes. Scenario-based testing of translation technologies with emergency personnel found domain-specific fixed-phrase tools preferred over general-purpose translation [7], a finding echoed in comparisons of speech-enabled phraselators [12]. Our phrase library is a much thinner artifact than any of these: a dispatcher-editable table of titles and descriptions with no clinical validation process behind it, whose machine translations are stored after first use and replayed without review.

Work on ASR in dispatch has been predominantly monolingual [13-15], as has downstream work on incident classification [16], severity prediction [17] and call-handler decision support [18]. Multilingual emergency communication systems have been engineered and evaluated on synthetic or laboratory material [19], and AI-enhanced prehospital networks have been evaluated for transport delay in single-language settings [28]. Against that background, the distinctive element here is the setting rather than the system. The technical artifact is an integration of commodity cloud services. What is uncommon is that it ran in a dispatch center with a highly multilingual caller population, including pilgrims and large resident communities, and left a record.

Data protection and clinical governance

Two governance findings bear on the service rather than on the analysis. The first, reported under Ethical considerations, is that 62.8% of the merged log passed through a consumer messaging platform. The second concerns the tool itself. Live call audio, which may contain names, addresses, and health information, is streamed over the public internet to Google Cloud Speech-to-Text at its "global" location without a regional endpoint, so processing was not restricted to Saudi Arabia. The speech services were used through an SRCA institutional account under Google's standard Cloud terms, but the translation step uses an undocumented endpoint without an API key that carries no contract, service level, or data-processing terms and can be rate-limited or withdrawn without notice.

Saudi Arabia's Personal Data Protection Law, fully enforceable since 14 September 2024, treats health data as sensitive and restricts its transfer outside the Kingdom [29]. No assessment against that law, and no formal information-security or clinical-safety review, was carried out before go-live, and use of the tool was agreed by center management only. These findings do not depend on the study's data, and they should be addressed before the tool remains in routine use.

Implications for practice and for the next study

For the service, several actions follow from this record and from the software audit, and none depends on further research. First, the stored translations of CPR and other critical phrases should be reviewed by fluent speakers of each target language before those phrases continue to be used, because the application replays a phrase's machine translation on every later use in that language and offers no way to inspect or correct it. Second, the tool needs a formal information-security, data-protection, and clinical-safety review, and the undocumented translation endpoint should be replaced with the documented Cloud Translation API under a service agreement. The remaining actions concern the record. The instruction-outcome field needs a written coding rule, because identical case descriptions are currently coded both ways. The second-action field needs to accept more than one action per call, because a single recorded action cannot confirm or refute the outcome code. The language field needs validation at entry, because 12 analyzed records contain a medical complaint or a generic case descriptor where a language should be, and one is blank. The date field needs to be mandatory and validated at entry, because 190 of 269 records (70.6%) could not be used for want of a valid date. And the use of a consumer messaging platform for operational case detail, which accounts for 62.8% of the merged log, should be reviewed and replaced with a supported record system.

For research, this study is best understood as establishing what is missing. A prospective evaluation of dispatch translation in this setting would need five things the present record lacks. Call-handling intervals would have to come from the computer-aided dispatch system, linked to the case log through the case number, rather than from staff recall. A sample of recorded calls, collected under an approved protocol, would have to be transcribed and rated by clinicians fluent in both languages, so that recognition and translation quality can be measured directly. Clinically consequential mistranslation would need its own analysis, because a confident wrong rendering of a symptom or an instruction is the safety concern that matters most. Per-language samples would have to be large enough to support per-language statements, which for every language in this record except Urdu means a much longer collection period or multi-center recruitment. Finally, the comparison a service actually faces, between AI translation, over-the-phone interpretation, and rostering bilingual dispatchers, would have to be made on accuracy, latency, and cost. A published protocol for a prospective observational study of over-the-phone interpretation in 911 calls defines call-level measures for the human-interpreter arm of such a comparison [30]. The present study supports none of these comparisons, and that gap is the reason to state the requirement.

Limitations

The limitations of this work are extensive and largely structural.

The study is retrospective and descriptive, with no control condition, no randomization, no pre-specified endpoint, and no adjustment for confounding. Nothing in it supports a causal claim.

The analyzed set is a minority of the record. Of 269 records in the merged log, 190 (70.6%) were excluded because they carried no valid date, and the exclusion removed the whole WhatsApp-derived stream, all records from four other regions, and eight of the 14 language labels, covering seven of the 13 distinct languages. The analyzed records therefore describe one center and a narrower language mix than the full log, and conclusions drawn from them may not generalize to the excluded records. The valid dates themselves could not be checked against the Arabic source. The only check available is that the recorded day of the week agrees with the date in the 20 analyzed records that carry both.

The study period was derived from the first and last valid record and was not defined in advance. It need not coincide with the period of deployment, and no valid record exists for any date before 19 May 2025 or after 14 October 2025. The log is also a set of cases that dispatchers chose to record, so it is not a register of all calls, and no call volume or rate can be derived from it. No temporal analysis is possible, and no part of the record can be attributed to the Hajj as distinct from year-round Umrah traffic.

Thirteen analyzed records carry no usable language label, and they supply 19.0% of the records coded as instructions given.

The record cannot distinguish pilgrims from resident callers, and much of the Urdu and Bengali demand may come from Saudi Arabia's large resident communities rather than from pilgrims.

The outcome field is an operational coding decision rather than a measurement. Its reliability is limited in two demonstrated ways: identical case descriptions coded both ways across 20 records, and a recorded rate that moves from 69.6% to 93.9% between adjacent date strata, a split that was itself chosen post hoc. Recording practice also changed in October 2025, when staff were asked to log every case and some records were entered retrospectively, so the two strata were not recorded in the same way. The second interpreter action holds one value per record, so it can confirm the field but can rarely test it. No alternative outcome measure can be constructed from these data, so the limitation cannot be addressed within this study.

The analysis was performed on an English version of an operational record kept in Arabic. That version was used as supplied, and no back-translation or other verification of it was carried out. In a report about translation, this is a material limitation, and it is not one the present materials can close.

Per-language sample sizes are too small for per-language conclusions. Only Urdu (44) reaches double figures among the analyzed records, and eight of the 14 language labels in the full log, covering seven distinct languages, have none at all.

The surveys are small (15 and 10 respondents), anonymous and therefore unlinkable, and conducted at a single center; both were administered after deployment. Their response rates of 100% rest on the authors' count of eligible dispatchers and could not be checked against individual invitations. Their administration windows overlap, the eligible groups are defined differently (all dispatch staff before deployment, only staff who had used the program afterward) and so differ in size, and the surveys are not assumed to be statistically independent. One author, who also developed the tool, wrote the items and distributed the forms, which may have made colleagues feel obliged to take part or to answer favorably. The retrospective survey of pre-deployment experience measures recall, not baseline, and the post-deployment improvement item is a retrospective self-comparison by dispatchers of the service that deployed the tool. The two instruments use incompatible bins on their shared understanding-time item. The retrospective survey form

presented one item twice, and one respondent answered the two copies differently. Both CPR items were answered by respondents who reported no relevant experience. Respondents rated Arabic output and could not assess source-language fidelity. The post-deployment accuracy item offered no option below "moderately accurate," and the improvement item offered only improvement options, so both are positively framed, and the improvement item is leading. No validation of the instruments is documented.

No objective measurement of the system exists: no word error rate, no translation quality score, no latency measurement, no assessment under controlled acoustic conditions. The translation step also depends on an undocumented web endpoint, so its behavior can change without notice, and a phrase translated once is replayed from local storage without review. None of the stored phrase translations, including those for CPR, has been checked by a speaker of the target language.

Both authors developed SMART TRANSLATION and work in the service that deployed it. This report describes the tool and its record and does not evaluate it, but the authors are not independent observers of it, and a reader should weigh the descriptive account with that in mind. Because the archived source tree carries version string 1.3.0 and the deployed builds are numbered 11 and 12, the statements about the application's code describe the archived source and are not claimed to hold identically for each deployed build. The go-live date of version 12 was not recorded, so no record can be assigned to a version.

Finally, ethics review followed data collection rather than preceding it, and no contemporaneous written participant information or consent documentation exists for the two surveys. The surveys were run as an internal service evaluation, and the IRB application disclosed that the data had already been collected.

Conclusions

A cloud-based speech translation tool was deployed on the dispatch consoles of a Saudi emergency medical center and was in documented operational use between 19 May and 14 October 2025. The 79 dated records analyzed describe the translator-assisted calls that dispatchers chose to document at one center, not the overall linguistic demand of the service, and the pilgrimage setting is context for the deployment rather than a finding about individual callers. The dominant documented use was establishing the caller's chief complaint. The recorded figure of 63 of 79 calls with instructions given is a documentation category, coded inconsistently, and not a measure of the tool's performance. These preliminary, setting-specific findings document deployment, operational use, and the ease of use reported by respondents. They do not establish translation accuracy, call-handling efficiency, patient safety, or patient benefit, and this report makes none of those claims.

The operational record was collected for service purposes and lacks the timing, linguistic, and clinical measurements that stronger claims would require, and its coding inconsistencies and missing dates limit the precision of even the descriptive findings. The value of this account is therefore twofold. It documents how a translation tool was used in practice, together with the governance gaps that accompanied its deployment, and it specifies what a credible prospective evaluation in this setting would need: standardized outcome definitions, call-handling intervals linked from the dispatch system, independent assessment of source-to-target translation fidelity, and a defined sampling strategy. Services deploying similar tools should build that measurement in from the start, because it cannot be recovered afterwards from an operational log.

Appendices

Appendix A. Pre-deployment dispatcher survey instrument

Administered anonymously in Arabic through Google Forms on 14 and 15 October 2025, after deployment, to the 15 eligible dispatchers of the SRCA Al-Madinah 997 center (15 responses). The survey asked dispatchers to recall the period before the tool was available. Each item allowed a single answer. The Arabic item and option wording is taken from the Arabic response export, and the English is the translation used in the analysis. The form presented Q5 twice, as two consecutive questions with identical wording; it is listed once. Where an item is introduced by "Response options," the list is complete. Where it is introduced by "Response options chosen by at least one respondent," only the options that respondents chose are known.

Q1. How often do you receive calls in a language other than Arabic or English?
كم مرة تتلقّى مكالمات بلغة غير العربية أو الإنجليزية
Response options: Often (كثيرًا); Sometimes (أحيانًا); Rarely (نادرًا); Never (لا أبدًا).

Q2. How did you handle these calls?
كيف كنت تتعامل مع هذه المكالمات
Response options: Try to communicate directly (أحاول التواصل مباشرة); Ask a colleague for help (أطلب المساعدة من زميل); Other (أخرى).

Q3. How clear was communication with the caller?
ما مدى وضوح التواصل مع المتصل

Response options chosen by at least one respondent: Very clear (واضح جدًا); Clear (واضح); Unclear (غير واضح); Very difficult (صعب جدًا).

Q4. How long does it usually take to understand the case?
كم من الوقت يستغرق عادة لفهم الحالة
Response options chosen by at least one respondent: Less than 2 minutes (أقل من دقيقتين); 2 to 5 minutes (من 2 إلى 5 دقائق); More than 5 minutes (أكثر من 5 دقائق).

Q5. Did the language barrier affect the speed of giving medical instructions?
هل كانت مشكلة اللغة تؤثر على سرعة إعطاء التعليمات الطبية
Response options chosen by at least one respondent: Yes (نعم); Sometimes (أحيانًا).

Q6. How many times have you tried to give CPR instructions to a caller who does not speak Arabic or English?
كم مرة حاولت إعطاء تعليمات الإنعاش القلبي الرئوي (CPR) لمتصل غير ناطق بالعربية أو الإنجليزية
Response options chosen by at least one respondent: Several times (عدة مرات); A few times (مرات قليلة); Once (مرة واحدة); Never (لم يحدث).

Q7. Did callers understand the CPR instructions correctly?
هل كان المتصلون يفهمون تعليمات الإنعاش القلبي الرئوي بشكل صحيح
Response options chosen by at least one respondent: Yes (نعم); Partially (جزئيًا); No (لا); Not sure (غير متأكد).

Appendix B. Post-deployment dispatcher survey instrument

Administered anonymously in Arabic through Google Forms on 15 and 16 October 2025 to the 10 eligible dispatchers of the SRCA Al-Madinah 997 center who had used the program (10 responses). Each item allowed a single answer. The Arabic item and option wording is taken from the Arabic response export, and the English is the translation used in the analysis.

Q1. How often do you receive calls in a language other than Arabic or English?
كم مرة تتلقّى مكالمات بلغة غير العربية أو الإنجليزية
Response options: Often (كثيرًا); Sometimes (أحيانًا); Rarely (نادرًا); Never (لا أبدًا).

Q2. How easy is the system to use?
ما مدى سهولة استخدام النظام
Response options: Very easy (سهل جدًا); Easy (سهل); Slightly difficult (صعب قليلًا); Difficult (صعب).

Q3. How accurate is the translation?
ما مدى دقة الترجمة
Response options chosen by at least one respondent: Very accurate (دقيقة جدًا); Accurate (دقيقة); Moderately accurate (متوسطة الدقة).

Q4. How clear are the translated instructions to the caller?
ما مدى وضوح التعليمات المترجمة للمتصل
Response options chosen by at least one respondent: Very clear (واضحة جدًا); Clear (واضحة); Confusing (مربكة).

Q5. How long does it take to understand the case now?
كم من الوقت يستغرق لفهم الحالة الآن
Response options chosen by at least one respondent: Less than 1 minute (أقل من دقيقة); 1 to 2 minutes (من 1 إلى 2 دقيقة); More than 2 minutes (أكثر من دقيقتين).

Q6. How many times have you given CPR instructions using the software?
كم مرة قدّمت تعليمات الإنعاش القلبي الرئوي باستخدام البرنامج
Response options chosen by at least one respondent: Several times (عدة مرات); A few times (مرات قليلة); Once (مرة واحدة); Never (لم يحدث).

Q7. How clear are the CPR instructions after translation?
ما مدى وضوح تعليمات الإنعاش القلبي الرئوي بعد الترجمة
Response options chosen by at least one respondent: Very clear (واضحة جدًا); Clear (واضحة); Unclear (غير واضحة).

Q8. How much improvement compared to before the software?
ما مدى التحسن مقارنة بما قبل البرنامج
Response options chosen by at least one respondent: Significant improvement (تحسن كبير); Slight improvement (تحسن بسيط).

Q9. Would you recommend using this system in real 997 operations?
هل توصي باستخدام هذا النظام في عمليات 997 الحقيقية
Response options chosen by at least one respondent: Yes (نعم); Maybe (ربما); Not sure (غير متأكد).

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Hameed Almarwani, Bader Alharbi

Acquisition, analysis, or interpretation of data: Hameed Almarwani, Bader Alharbi

Drafting of the manuscript: Hameed Almarwani, Bader Alharbi

Critical review of the manuscript for important intellectual content: Hameed Almarwani, Bader Alharbi

Disclosures

Human subjects: Informed consent for treatment and open access publication was obtained or waived by all participants in this study. Saudi Red Crescent Authority Institutional Review Board (SRCA IRB) issued approval IRB-25-0042S. The Saudi Red Crescent Authority (SRCA) Institutional Review Board (registered with the Saudi National Committee of Bioethics under number H-01-R-110) reviewed this study and approved it in the exempt category on 30 December 2025 (IRB log number IRB-25-0042S). The application was submitted on 3 November 2025, after the operational records and the survey responses had been collected, so the board's determination was made on data that had already been collected. The study analyzed retrospective operational call records containing no patient identifiers, together with anonymous, voluntary staff surveys. No patient was contacted for the study. No signed consent forms are held. Participation in the surveys was voluntary and consent was inferred from completion of an anonymous form. The SRCA Research and Studies Administration replied in writing on 29 September 2026 to the authors' notification of approval to publish. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** Both authors are employed by the Saudi Red Crescent Authority (SRCA), the service that operates the dispatch center described in this report and that deployed SMART TRANSLATION. Both authors developed SMART TRANSLATION, so they report on a tool they built and on a service in which they work. One author (HA) wrote the survey items. The authors have no financial interest in the tool and no relationship with Google LLC, the provider of the cloud services used by the tool and of the survey platform, other than as users of those services. .

Acknowledgements

The authors thank the dispatchers and operational staff of the Saudi Red Crescent Authority (SRCA) Al-Madinah 997 emergency call center, who used the tool during one of the busiest periods of the operational year and who completed the surveys reported here. Use of artificial intelligence: Claude Sonnet 5.5 (Anthropic), accessed through Claude Code, was used only to write and debug the Python scripts used in this work and to improve the language and clarity of the manuscript. It was not used to collect, analyze, or interpret the data, generate scientific results, or make research decisions. The authors reviewed all AI-assisted output and take full responsibility for the content of the article.

References

1. Meischke H, Chavez D, Bradley S, Rea T, Eisenberg M: Emergency communications with limited-English-proficiency populations. Prehosp Emerg Care. 2010, 14:265-71. 10.3109/10903120903524948
2. Nuño T, Bobrow BJ, Rogge-Miller KA, et al.: Disparities in telephone CPR access and timing during out-of-hospital cardiac arrest. Resuscitation. 2017, 115:11-6. 10.1016/j.resuscitation.2017.03.028
3. Del Rios M, Bartos JA, Panchal AR, et al.: Part 1: executive summary: 2025 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation. 2025, 152:S284-312. 10.1161/CIR.0000000000001372
4. Sanko S, Feng S, Lane C, Eckstein M: Comparison of emergency medical dispatch systems for performance of telecommunicator-assisted cardiopulmonary resuscitation among 9-1-1 callers with limited English proficiency. JAMA Netw Open. 2021, 4:e216827. 10.1001/jamanetworkopen.2021.6827
5. General Authority for Statistics: Hajj statistics 1446H/2025 (executive summary) . (2025). Accessed: October

1, 2026: https://www.stats.gov.sa/documents/d/guest/hajj-report-1446h-en-1-.
6. Alrabie T, Brown M, Rice B, Marsh L: Disaster preparedness and response among healthcare professionals during the Hajj: a systematic literature review. Healthcare (Basel). 2025, 13:1571. 10.3390/healthcare13131571
7. Turner AM, Choi YK, Dew K, et al.: Evaluating the usefulness of translation technologies for emergency response communication: a scenario-based study. JMIR Public Health Surveill. 2019, 5:e11171. 10.2196/11171
8. Noack EM, Schulze J, Müller F: Designing an app to overcome language barriers in the delivery of emergency medical services: participatory development process. JMIR Mhealth Uhealth. 2021, 9:e21586. 10.2196/21586
9. Müller F, Hummers E, Schulze J, Noack EM: Usability of an app to overcome language barriers in paramedic care (Article in German). Notf Rett Med. 2022, 25:570-7. 10.1007/s10049-021-00913-w
10. Müller F, Schröder D, Noack EM: Overcoming language barriers in paramedic care with an app designed to improve communication with foreign-language patients: nonrandomized controlled pilot study. JMIR Form Res. 2023, 7:e43255. 10.2196/43255
11. Müller F, Schröder D, Noack EM: Correction: overcoming language barriers in paramedic care with an app designed to improve communication with foreign-language patients: nonrandomized controlled pilot study. JMIR Form Res. 2025, 9:e83353. 10.2196/83353
12. Boujon V, Bouillon P, Spechbach H, Gerlach J, Strasly I: Can speech-enabled phraselators improve healthcare accessibility? A case study comparing BabelDr with MediBabble for anamnesis in emergency settings. Proceedings of the 1st Swiss Conference on Barrier-free Communication: Methods and Products. 2018, 39:51.
13. Valizada A, Akhundova N, Rustamov S: Development of speech recognition systems in emergency call centers. Symmetry. 2021, 13:634. 10.3390/sym13040634
14. Ungureanu D, Toma SA, Filip ID, et al.: ODIN112-AI-assisted emergency services in Romania. Appl Sci. 2023, 13:639. 10.3390/app13010639
15. Minulescu DE, Toma SA: Whisper based speech recognition for emergency services. 2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, Piscataway, NJ. 2025, 1:6. 10.1109/ECAI65401.2025.11095458
16. Ferri P, Sáez C, Félix-De Castro A, Juan-Albarracín J, Blanes-Selva V, Sánchez-Cuesta P, García-Gómez JM: Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch. Artif Intell Med. 2021, 117:102088. 10.1016/j.artmed.2021.102088
17. Abi Kanaan M, Couchot JF, Guyeux C, Laiymani D, Atechian T, Darazi R: Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction. Array. 2024, 21:100333. 10.1016/j.array.2023.100333
18. Attiah A, Kalkatawi M: AI-powered smart emergency services support for 9-1-1 call handlers using textual features and SVM model for digital health optimization. Front Big Data. 2025, 8:1594062. 10.3389/fdata.2025.1594062
19. Rafi RA, Ahmed S, Venkateshperumal D, Khokhar A, Arifuzzaman M, Azad AK, Alyami SA: Multilingual low-latency emergency VoIP system using LLM for speech reconstruction and blockchain for secure data archiving. IEEE Access. 2025, 13:150101-22. 10.1109/ACCESS.2025.3600364
20. Moser D, Stanic N, Sariyar M: Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions. JAMIA Open. 2025, 8:ooaf147. 10.1093/jamiaopen/ooaf147
21. Luo X, Zhou L, Adelgais K, Zhang Z: Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: a comparative study of four AI-powered engines. J Healthc Inform Res. 2025, 9:494-512. 10.1007/s41666-025-00193-w
22. Pinnock H, Barwick M, Carpenter CR, et al.: Standards for Reporting Implementation Studies (StaRI) statement. BMJ. 2017, 356:i6795. 10.1136/bmj.i6795
23. Eysenbach G: Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res. 2004, 6:e34. 10.2196/jmir.6.3.e34
24. General Authority for Statistics: Saudi Census Statistics 2022: population statistics . (2023). Accessed: November 14, 2026: https://www.stats.gov.sa/en/statistics-tabs?tab=436332&category=417653.
25. Google Cloud: Cloud speech-to-text documentation. (2026). Accessed: October 1, 2026: https://cloud.google.com/speech-to-text/docs.
26. Google Cloud: Cloud text-to-speech documentation. (2026). Accessed: October 1, 2026: https://cloud.google.com/text-to-speech/docs.
27. Hunter JD: Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007, 9:90-5. 10.1109/MCSE.2007.55
28. Kim JH, Kim MJ, Kim HC, Kim HY, Sung JM, Chang HJ: A novel artificial intelligence-enhanced digital network for prehospital emergency support: community intervention study. J Med Internet Res. 2025, 27:e58177. 10.2196/58177
29. Saudi Data and Artificial Intelligence Authority: Personal Data Protection Law, issued by Royal Decree No. M/19 of 9/2/1443H (16 September 2021) and amended by Royal Decree No. M/148 (27 March 2023). (2023). Accessed: November 8, 2026: https://sdaia.gov.sa.
30. Meischke H, Fogel CA, Garrison MM, et al.: Protocol for an observational prospective study on the use and impact of over-the-phone interpretation in 911 calls with language discordance. BMC Public Health. 2025, 25:3413. 10.1186/s12889-025-24538-7