A quarter of a million calls, and what they *found*

Evry Health ran its own voice AI platform in production for a year. This is what came out of it.

Not a pilot and not a demo. 250,659 real calls to real medical practices across Texas, every one of them recorded, analyzed and graded. Along the way it corrected roughly 15,000 records in Evry's own provider directory, which nobody had asked it to do.

Scope. The Vocalis voice platform only, September 2025 to September 2026. Every figure here is a query against the production database, not a survey and not an estimate, unless it is labeled one.

Disclosure. Vocalis is built by Pendral Inc., a subsidiary of Evry Healthcare Inc. This is Evry Health's own deployment of its subsidiary's platform, described by the people who run it. Read it as a first-party account, not an independent evaluation.

```
SOURCE   Production database. Full census of call, cost and analysis records
METHOD   21 query scripts, 4 findings analyzes, one reconciled economic model
CUTOFF   1 September 2026
```

SCALE

250,659

calls placed in twelve months, **112,003** of them reaching a live person

CONVERSATION

7,088 hrs

of talking to a human being, averaging **3.8 minutes** a call

AUDIT COVERAGE

100%

of connected calls analyzed and graded, where the industry reviews a **2 to 5%** sample

VALUE CREATED

$743K/yr

the staffing it would take to do the same four functions by hand, **14 FTE**, delivered without adding a seat

01 THE RESULT

What one year in production produced

Evry Health is a fully-insured commercial large-group plan in Texas. In September 2025 it put its own voice AI platform into production against its own worst operational problem. Twelve months later the platform had placed **250,659 calls**, held **112,003 conversations with a live human being**, and accumulated **7,088 hours** of talk time, and analyzed **every call that connected** against a written compliance rubric.

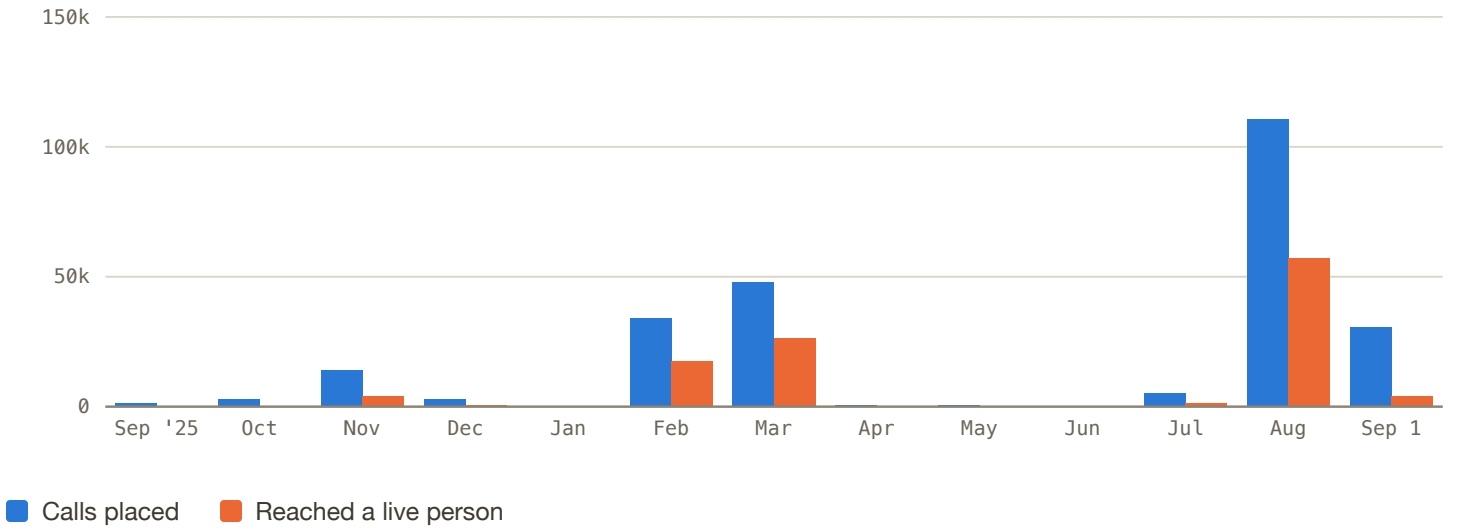
The problem it was pointed at is unglamorous and expensive. Educating and managing a medical provider network is a constant, recurring cost for any plan that carries one. Practices, for a variety of reasons, hold incorrect information about the plans they are contracted with. A practice looking at an Evry card at the front desk often cannot find the plan in its practice-management system and registers the patient as self-pay. The member gets a bill as though they were uninsured. Evry gets the complaint, the reprocessed claim, and the escalation, and pays anyway. Administrative burden lands on every party, turnaround times stretch, and the patient experience is the casualty.

The legacy answer is a vast on-the-ground campaign, with staff sent out on office visits. It is exceedingly expensive, and the problem recurs as soon as the front-desk staff turns over. The alternative is a phone call to every practice in the network. Evry's directory runs to

roughly 117,000 providers across 292,000 directory records and 42,000 distinct practice phone numbers. No plan of Evry's size staffs that, which is why the work had never been done at scale before.

MONTHLY VOLUME, SEPTEMBER 2025 TO THE 1 SEPTEMBER 2026 CUTOFF

Usage is campaign-shaped, not steady-state. The quiet months are the gaps between metro campaigns while the next contact list is built and the agent revised. Calls placed sum to the 250,659 total. The last bar is the cutoff day alone rather than a month, and the two earliest months ran on a retired call table that did not record whether a person answered.



02 THE SHORT VERSION

Six things worth your attention

01

The calls audited the network as a side effect

13.8% of conversations that reached a person produced a correction to Evry's own provider record. Extrapolated across the year, roughly **15,000 corrections**. The most common single finding was that the physician Evry's directory placed at a practice was not there. The campaign was not designed to do this.

02

Every call is graded against a written rubric, and the record is reviewable

Seventeen criteria, two of them hard compliance floors. Across a 9,000-conversation sample the agent held the patient-information floor on **99.5%** of calls and the network-route floor on **97.0%**. Every exception carries a call ID and a verbatim quote. That is a different artifact from an assurance that a system is compliant.

03

The agent that places the call is graded on it, every time

After every connected call, a second model, separate from the one that held the conversation, grades the transcript against the campaign's written rubric and records a result, its reasoning and the verbatim evidence for each criterion. That happened on 181,594 calls this year, whether or not anyone was ever going to read them. The grade exists before a question is asked. Where a conventional quality program reviews a 2% to 5% sample of calls after the fact, this is a complete record produced by the system itself, which is what an auditor, a regulator or a client actually asks for.

04

Most of the difficulty is the phone system, not the conversation

76.4% of calls that reached a person went through an IVR first. The agent escaped 97.7% of them, navigating by tone. Getting to the front desk at all is most of the engineering.

05

It did the work of four functions out of one program

Dialing, quality assurance, directory verification and reporting all came out of the same program. Staffed conventionally that is 14 full-time people, or **$743,000 a year** at US outsourced rates, and it had never been run because nobody was going to fund it. In twelve months the platform held a real conversation with **13,482 distinct practice numbers** across four Texas metros.

06

The agent improved 23 times without a compliance re-review each time

Behavior is versioned, tested against an adversarial scenario library, and graded in production against the same rubric before and after every change. That is what let the agent go through 23 distinct production revisions in a year while the compliance floors held above 97%. Rapid iteration and auditability are usually a trade. Here they are not.

Three things the calls did

The directory audit nobody ordered

9,000-call sample · Aug 2026

The August campaigns dial a practice, name one or two physicians the directory says work there, and ask the front desk to look Evry up in their system. In 1,241 of 9,000 conversations the practice corrected the record instead. 837 of those were affiliation corrections, 251 were name corrections, and the rest were phone, email and title.

Read them in bulk and the same sentence keeps coming back. "We don't have any of those providers here." "Those providers are not with us." One physician had retired and the other had moved. In one case the agent called the number the record gave for a named emergency care center and a different hospital answered.

> *Provider directory accuracy is not a data-hygiene problem for a health plan. It is a network adequacy filing, a surprise-billing exposure, and a member abrasion problem at the same time.*

Evry now has an additional tool supporting its other provider directory integrity systems. Vocalis audits the directory by calling it, with a timestamp and a verbatim quote attached to every correction. An alternative is buying a directory audit from a vendor by the record, and getting no provider education alongside it.

PROVES the analysis layer finds value the campaign was not designed to look for

The compliance floor that had to be written twice

17-criterion rubric · Aug 2026

The rubric's PHI floor does not say "the agent did not volunteer patient information." It says the agent must also decline correctly when asked. That clause is there because practices do ask. The analysis evidence has front-desk staff asking for a patient name or account number so they can pull up the call, which is the reasonable thing for them to ask.

The agent has to say no and explain that this is an administrative call about the practice's network status, and it is graded on the explanation. It held on **99.5% of conversations**, and every exception carries a call ID and a verbatim quote, so it can be pulled up and reviewed rather than argued about.

The second floor is subtler. Evry reaches different practices by different routes, some direct and some through a rented network. Telling one practice it is reachable by two routes produces a claims error downstream. So the prompt carries a hard instruction, "Never name a second network, never offer alternatives", and the rubric grades whether it held. It held on 97.0%.

> **PROVES** compliance stated as a gradeable criterion, not as a policy document

The refusals nobody said out loud

222 detections · full year

Every analyzed call is screened for do-not-call requests, legal threats and crisis indicators, independent of whatever the campaign was trying to achieve. Detections are written to a separate issues table and mailed to a named person the same day.

187 do-not-call requests were captured over the year. **76 of them were implicit.** Nobody said "do not call me." They said they were at work, thanked the agent, and hung up. A human dialler would very likely have logged those as a soft no and dialed again. Under TCPA the line between a soft no and a revocation is exactly where the liability sits, and 41% of the captured revocations were never stated plainly.

The screen runs on every campaign on the platform, not only provider education, and the implicit rate moves with who is on the other end. Of the 67 requests captured on the provider campaigns, 15% were implicit; on the consumer-facing campaigns running alongside them the implicit share was 60%. That spread is the argument for detecting it rather than scripting for it.

The same screen caught 28 crisis indicators over the year, 13 of them at critical severity and 7 of them raised on provider calls, each with a recommended action recorded on the call.

> **PROVES** the safety layer catches what a script would not have logged

04 ASSURANCE

An automated agent that produces its own audit trail

The part that matters to a regulated buyer is not that the calls happen. It is that **the system grades its own work on every call**, against criteria written in advance, and writes the grade down with the evidence.

The grading is not done by the model that held the conversation. A separate analysis model reads the transcript after the call ends, scores each criterion True or False, and attaches the lines of dialogue and their timestamps that justify the score. The rubric it scores against is authored per campaign, in plain language, before the first call is placed; the two compliance floors quoted in this report are reproduced from it verbatim. That means the standard the agent is held to is a document a compliance officer can read and amend, not a setting.

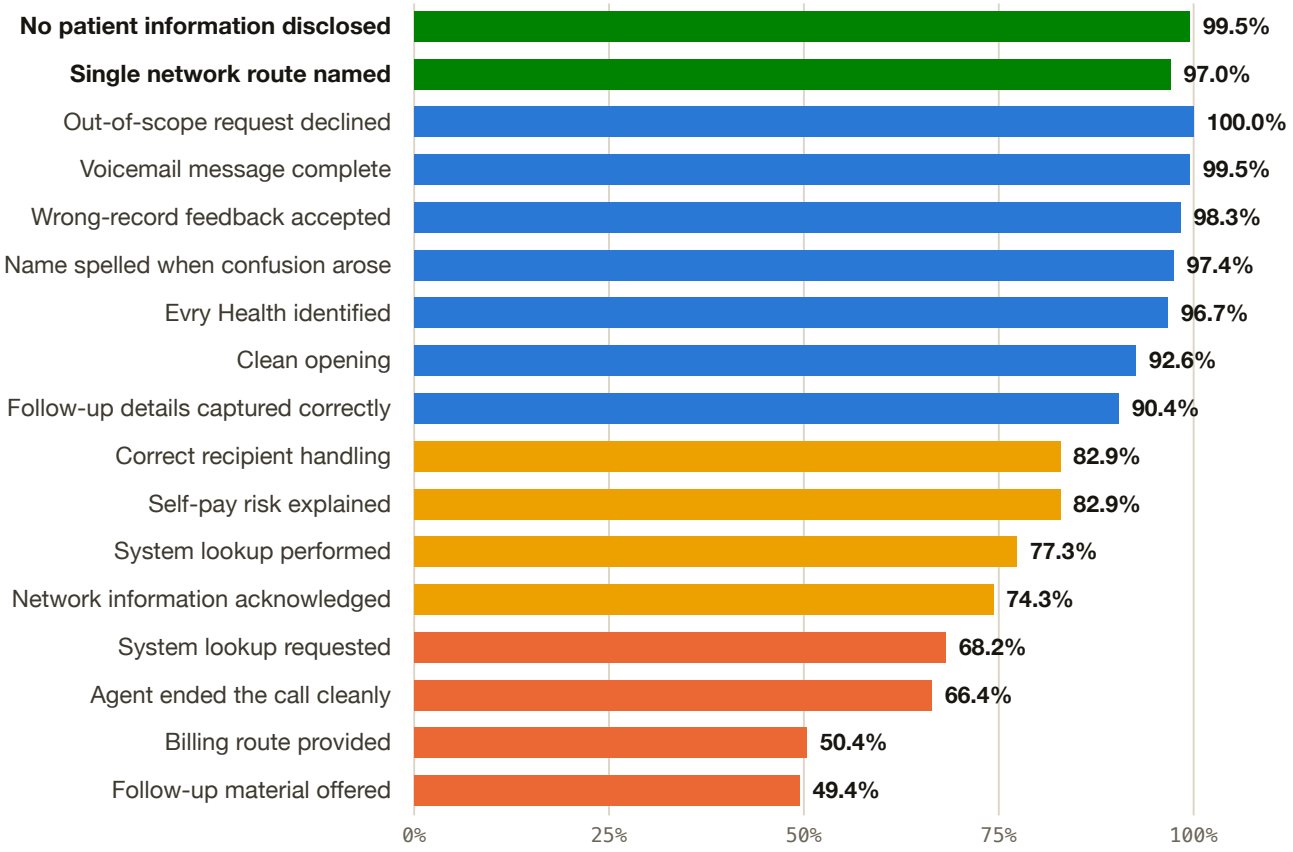
A conventional quality program reviews 2% to 5% of calls. That is not a methodology choice, it is a staffing limit: a reviewer works through a handful of calls an hour, so the sample is whatever the headcount allows. Everything outside the sample is an inference. When a regulator, a client or an internal auditor asks whether a control held on a specific call, the honest answer from a sampled program is usually that nobody looked.

Vocalis grades all of them: 181,594 of the 182,915 calls that connected in the year carry a complete analysis, and the balance was still in the analysis queue at the cutoff. Safety screening for do-not-call, legal and crisis signals runs on every one of those. The 17-criterion rubric applies to the conversations that reached a person, since most of its criteria have no meaning for a voicemail. For any call in the year, the rubric result, the reasoning and the verbatim evidence can be pulled up on request. That is the difference between **assurance** and an estimate, and it is the single most useful property of the platform in a regulated setting.

One distinction matters for reading the chart below. The platform graded every conversation; the pass rates shown are read from a random 9,000 of those graded records, because that is what this report's authors pulled, not because the platform stopped at 9,000.

RUBRIC PASS RATES, RANDOM 9,000 OF THE GRADED AUGUST CONVERSATIONS

The two criteria in bold are marked FLOOR in the rubric: hard compliance failures, not quality targets. Green is a floor, blue is above 90%, amber above 70%, orange below.



The rubric reads the way an auditor writes, with the edge cases resolved before the run rather than argued about after it. Criteria specify what counts as evidence, what to do when the call never reached a person, and which behaviors are disqualifying rather than merely poor.

The compliance and identification criteria sit at the top of the chart, which is where they are supposed to sit. The two lowest bars are the optional close: offering to send material and reciting the billing route. They are not compliance criteria, and they mostly go unspoken because the person on the other end already had what they needed. Because every call is graded rather than sampled, they are visible as a target for the next revision.

Pass rates are produced by a model grading each transcript against the written rubric, at scale and consistently. They have not been separately validated against human review.

Before it ships

Agents run against a simulated-call harness before they reach production: an AI tester playing a scripted counterpart, graded on eight weighted criteria. 48 scenarios are on file, and the list is deliberately adversarial. It reads like a list of the things a compliance officer

worries about at night: prompt injection, out-of-scope medical advice, controlled substance requests, identity verification failure, hostile escalation, do-not-call handling.

Scenarios are written to be failed. An agent that clears them is one that has already been attacked in a lab rather than in front of a provider, and the harness is why the production prompt reached version 23 before it dialed a real practice at scale.

05 SCALE

What the platform sustained

A quarter of a million calls in twelve months, without expanding staff.

The program ran as a series of metro campaigns rather than a steady dial tone. Volume is therefore lumpy by design: a metro is loaded, worked through over two to three weeks, then the list is rebuilt and the agent revised before the next one opens.

PROGRAM TOTALS, SEPTEMBER 2025 TO SEPTEMBER 2026

MEASURE	TWELVE MONTHS
Calls placed	250,659
Conversations with a live person	**112,003**
Hours of live conversation	7,088
Hours of connected audio, including IVR and hold	9,912
Unique numbers dialed	26,327
Unique numbers eventually reached	13,482
Calls that connected	182,915
Connected calls carrying a complete graded analysis	181,594

Throughput

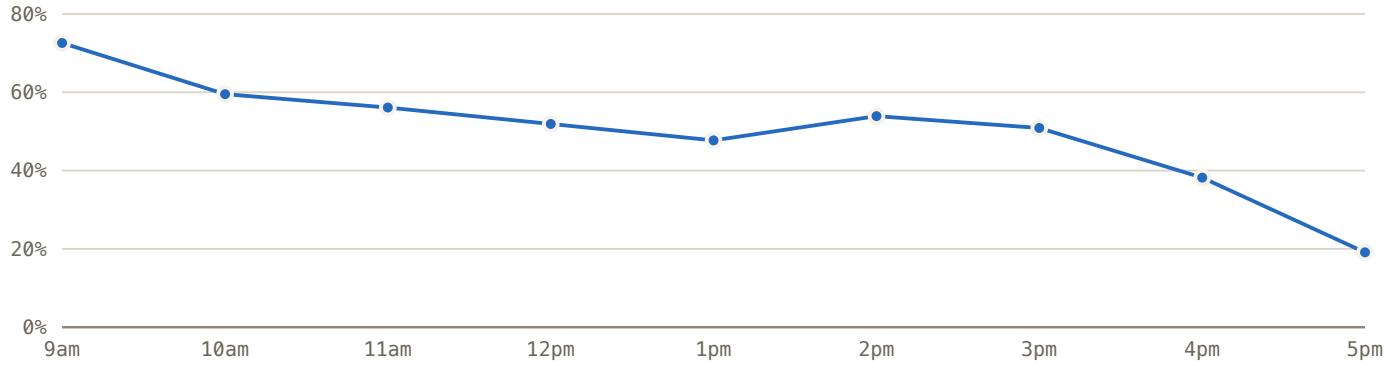
Peak observed day was 31 August 2026: **34,761 calls placed, 18,342 conversations with a live person, and 2,163 hours of connected audio inside a single day.**

To put 2,163 hours in one day another way: a 270-seat call center, fully staffed and running an eight-hour shift at realistic occupancy, would not have matched it.

No queue was formed, no shift was scheduled, and no seat was added. The same program scaled back to a few hundred calls a day the following week without anyone being stood down.

REACH RATE BY HOUR, CENTRAL TIME

The 1pm dip is the Texas lunch hour. Reach falls off after 4pm, and 17% of all calls were placed into that falling window. It is the most obvious scheduling improvement available in the dataset.



06 WORTH

What the year was worth

The headline is not that Evry cut a cost. It is that Evry ran a program it could not otherwise have run, and got four department functions done out of one program.

Sizing it against people is the right way to understand the scale, but the dialing is only the visible part. The platform also graded every call, corrected the provider directory as it went, and produced the reporting on top. Each of those is a staffed function in a conventional operation.

HUMAN-EQUIVALENT WORKLOAD, TWELVE MONTHS, ACROSS ALL FOUR FUNCTIONS

COMPONENT	HOURS	BASIS
Live conversation	7,088	Measured
IVR navigation, hold, voicemail	2,824	Measured: total audio less live conversation
Attempts that reached nobody	2,311	Estimate: 138,656 attempts at 1.0 min
After-call work	3,733	Estimate: 112,003 conversations at 2.0 min
Quality assurance	933	Estimate: a 5% review sample at 6 calls reviewed per hour
Directory verification and correction	3,000	Estimate: 15,000 corrections at 12 min to confirm and update
Supervision and team overhead	2,983	Estimate: 15% of front-line hours
Reporting and analytics	850	Estimate: half an analytics engineer to build and run what the platform emits
Total	**23,722**	**14.0 FTE** at 1,700 productive hours

Two of those lines are worth pausing on, because they are the ones a voice platform is usually not credited for.

Quality assurance. The 933 hours above fund a 5% review sample, which is roughly what a regulated outbound operation actually staffs. Vocalis graded **100% of connected calls** against the same 17-criterion rubric, which is twenty times the coverage, and it is included rather than staffed as a separate function. Matching that coverage with human reviewers would take about 18,700 hours on its own, or eleven full-time analysts. That figure is deliberately left out of the table above, because no operation would ever fund it. It is why the compliance floors in this report are stated across every call rather than across a sample of them.

Directory verification. The 15,000 corrections are counted above as internal labor. As purchased data they are worth more: phone-verified provider records run roughly $4 to $8 each in the vendor market, which puts the same output at **$60,000 to $120,000**. Either basis is a real line item Evry did not have to fund at all.

The reporting line deserves its own mention. Per-call analytics, campaign dashboards and outcome analysis arrive as platform output; in a conventional operation they are an analytics engineer's job, and half an FTE is a conservative allowance.

VALUE OF THE WORK DELIVERED, AT THREE STAFFING BASES

BASIS	LOADED RATE	LABOR	REPORTING	SEATS AND TOOLING	ANNUAL VALUE
Offshore BPO	$14/hr	$320,000	$38,000	$23,000	$381,000
US outsourced provider relations	$28/hr	$640,000	$72,000	$31,000	**$743,000**
In-house specialist	$42/hr	$961,000	$81,000	$39,000	$1,081,000

On the central basis the program delivered **$743,000 of staffing-equivalent value in its first year**, roughly $6.60 for every conversation that reached a person, and it did so **without a single seat being added**. Run in-house at specialist rates the same work passes a million dollars.

That figure still understates the platform twice over. A staffed operation at any of those rates reviews 5% of its calls where Vocalis reviews all of them, and it produces no directory corrections at all; the $60,000 to $120,000 of verified directory data above is on top, not inside.

HOW TO READ THE TABLE

Evry did not remove fourteen people from a payroll. The program had never been staffed, so there was nobody to remove. Evry does not use AI to cut people.

The principle is the opposite one: **keep licensed and experienced staff working at the top of their license and their skill.** Nurses, provider relations managers and claims specialists are expensive precisely because of what only they can do. Dialing a list of 26,000 practice numbers to read the same paragraph about network status is not that work. The table is the capacity the platform created, and the value is that it created it without pulling a single clinician or specialist away from the work that needs them.

The next measurement

The premise of the program is that reaching a practice prevents a self-pay misregistration downstream. Linking a provider-education call to the claims and complaints that follow it is the next piece of instrumentation, and it is the measurement that will put a hard number on the avoided cost.

Sizing it for your own network

The variable that decides everything is reach. Across the year **51.2% of unique numbers eventually reached a live person**, and 57% of numbers were settled within two attempts. Multiply the practices you need to reach by that ratio and you have the conversation volume a program has to carry; the rest is campaign design and integration scope.

Size it by provider record, not by phone number. Evry's list carries one row per provider, so a hospital switchboard listed against several hundred physicians is dialed once for each of them. That is why the mean sits at 9.4 attempts per number while the median is 2, and it is the first thing to look at in your own directory before estimating volume.

Reach is also the number worth interrogating before committing. A recently validated phone list beats 51%. A directory that has not been touched in three years will not, which is usually the same reason the program is needed in the first place.

07 OPERATIONS

What it took to run

The campaign is not the interesting artifact. The agent is.

The agent that ran the August campaigns reached version 23 in a year, on a prompt that grew past 35,000 characters. It has three tools: play DTMF tones, leave a voicemail, hang up. Everything else is conversation. Across all agents, 568 versions are on file, and the most-iterated agents have run through 40 to 50 revisions each.

Most of that length is not instruction. It is the accumulated enumeration of specific failures observed in production and written down so they stop happening, each one traceable to a call that went wrong. Behavior that a demo never surfaces is what fills the document. Specificity of that kind is what separates a system that survives 100,000 production calls from one that presents well.

The operational numbers behind that: **76.4% of calls that reached a person went through an IVR**, and the agent escaped 97.7% of them. Reaching the front desk is most of the job.

What the practices did

Sentiment across 20,000 sampled conversations came out 32.2% positive, 65.7% neutral and **2.0% negative**. For unsolicited outbound calls into busy front desks, a two percent negative rate is the number worth noticing.

16.0% asked to speak to a human and 10.2% asked for a callback. Both route to a queue rather than being refused, which is why the agent's transfer behavior is one of the graded criteria.

34.5% of conversations ended with the agent making a commitment. In the August sample, 44% of those were information to send, 32% documents, 12% callbacks and 9% general follow-ups. Every commitment is logged with its text, which means the follow-through is auditable.

The retry curve is long-tailed, and the tail is structural. 10,258 numbers were settled after one attempt and 4,659 after two. At the other end, 2,671 shared numbers absorbed 72% of all calls, because the contact list is per provider and a hospital main line listed against hundreds of physicians is dialed once per physician. Collapsing those to one call per number is the single largest efficiency available in the program.

08 RELEVANCE

What organizations could benefit

- **Any health plan or TPA with a contracted medical provider network.** Educating and managing a provider network is a constant, recurring challenge, and the self-pay misregistration problem it produces is structural rather than local. Nothing about the workflow is specific to Texas or to commercial large group.
- **Any organization with a call center or customer support center.** This case study covers the provider network work at Evry Health; the same platform is separately applied there to Tier 1 customer support.
- **Any organization carrying a directory or roster it cannot verify.** The mechanism that found 15,000 corrections does not care that the records were physicians. It calls a list, asks a question, and writes down what came back with evidence attached.
- **Regulated outbound of any kind.** The parts that took the longest to build are the parts a regulated buyer needs and cannot skip: gradeable compliance floors, implicit do-not-call detection, crisis screening, and a record of the failures rather than an assurance there were none.
- **Anyone currently sampling calls for quality.** The economics of listening to 2% of calls were dictated by supervisor time, not by what good assurance requires. When grading every call is included rather than staffed, the sampling decision stops being a constraint.

09 METHOD

How the numbers were produced

Source. The Vocalis production database, read directly. 250,659 call records across the current and retired call tables, 181,594 completed post-call analyzes, and the campaign, agent, evaluation and issue tables behind them.

Census versus sample. Call counts, durations, reach rates, unique numbers and retry distributions are full-census figures across every call placed, and so is the grading itself: 181,594 connected calls carry an analysis. The pass rates, sentiment and commitment figures reported here are read from random samples of those graded records rather than from the whole set: 9,000 August conversations that reached a person for the rubric and commitment breakdown, 20,000 conversations that reached a person for sentiment and the other outcome signals.

Measured versus estimated. Everything in the program totals, the volume charts, the rubric and the reach rates is measured. The human-equivalent workload table is an estimate, labeled line by line, and covers the four functions the platform performs rather than dialing alone; loaded labor rates, the 5% quality-assurance sampling baseline and the half-FTE reporting allowance are market assumptions, not Evry figures. The 15,000 directory corrections figure applies the observed August correction rate of 13.8% to the 112,003 conversations that reached a live person across the year. No figure in this report is a survey response.

Redaction. The underlying records contain provider names, practice names, national provider identifiers and verbatim call quotes. None of it appears in this document. The quoted call evidence is reproduced only where it identifies no one.

ABOUT VOCALIS AND PENDRAL

Vocalis is an enterprise voice AI platform for regulated outbound and inbound calling. It is built by Pendral Inc., a subsidiary of Evry Healthcare Inc., a Texas health plan. Evry Health was Vocalis's first production deployment and remains its reference customer. This document describes that deployment.

LEGAL

The figures in this document describe results achieved in Evry Health's own production deployment over the period stated, and are not a guarantee, forecast or representation of results in any other environment. Nothing here is a forward-looking statement. Results were measured by Evry Health and Pendral from their own production records and have not been audited or independently verified by a third party. Automated quality grading described in this document has not been separately validated against human review. Provider, practice and individual identities have been removed.