CYARA

WHITEPAPER

WebRTC in Healthcare: From Risk to Scale

De-risking real time communication (RTC) in enterprise networks: validate the path, assure quality, observe live reality, and prove capacity

Executive summary

- **Reality:** WebRTC is no longer just for chats. Instead, this critical infrastructure powers virtual visits, remote monitoring, imaging reviews, and clinical trials. Enterprise networks are private and policy-heavy—designed for security, not real-time media—so day-to-day looks messy. Speed tests pass while production lags at peak. Regions drift after failover. One room or SSID is always worse. Long sessions drop at repeatable minute marks. These aren't edge cases; this is what clinicians feel at their busiest hours. Microsoft's own guidance tells customers to avoid hairpinning cloud traffic for latency reasons, which confirms how widespread this is.
- **Risk:** Small packet-level issues become longer calls, repeats, missed SLAs, and poorer customer experience (CX) and agent experience (AX). These failures often appear under sustained, concurrent use or right after a change window.
- **The solution:** You need constant visibility into how your systems are actually performing, under real-world conditions. With the right WebRTC testing and monitoring solutions, you can validate system performance and scalability under soak, identify real-world drift, and gain key insights into real-time quality.
- **End result:** This key investment into your CX strategy gives you a network you can test and observe from end to end. Fewer escalations. Faster, safer rollouts. Clear compliance-routing. Measurable voice and video quality at busy hour, with the power to drill straight to root cause.

The CTO's 30-second reality check

Speed tests are liars: A 500 Mbps connection means nothing if the path has 2% packet loss or jitter spikes during imaging syncs.

Security is the bottleneck: Your Zscaler, VPNs, and Firewalls are designed to inspect and delay. WebRTC needs to bypass and flow.

The "Monday at 10 AM" rule: If it doesn't work when the hospital is full, it doesn't work. Only soak tests can catch concurrency limits.

Wi-Fi is a living organism: A cart in Room 512 works today but fails tomorrow because of sticky clients and channel interference. You must continuously monitor your systems.

The cost of silence: When calls fail, clinicians don't open tickets. Instead, they switch to FaceTime or personal phones, compromising compliance.

CYARA

Introduction

It's 10 a.m. on a Monday at your hospital. Clinics are full, imaging is syncing, backups have just started. A remote consultation is about to begin. The app looks perfect in the lab. On this network, right now, will the call set up, stay stable, and meet policy?

In controlled tests, performance looks great. On real enterprise paths, network choices change outcomes at peak. The same application that behaves flawlessly at home can struggle in production when the network is busy. That hurts care delivery, trust, and cost.

Hospital networks are private, controlled, and heavily audited: VPNs, traffic backhauling, tight egress rules, proxies, layered network address translation (NAT), and Wi-Fi that must serve carts and rooms across the campus. Compliance architectures (VPNs, proxy PACs) are often the primary enemy of real-time media quality.

Whenever you connect patients and clinicians over WebRTC, you need confidence that the system will work every time and that the network path is actually enabled for real-time media. This pattern is not limited to hospitals. You see the same issues in insurance claims centers, banking contact centers, university helpdesks, retail town halls, and global business process outsourcing (BPO) sites. The logo changes. The symptoms do not.

Local ping or iperf can pass while voice breaks. Ten percent ping loss is annoying; 1-2% sustained media loss is brutal. Trust end-to-end WebRTC metrics. With the right testing and monitoring in place, you can spot the real bottlenecks early and deliver the quality patients and staff expect.

Note: While the examples here are hospital-centric, the same network patterns appear in finance, insurance, and large retail. Software-defined wide area networks (SD-WAN) backhaul, policy hubs, and split domain name system (DNS) don't care what the logo is; the symptoms are the same at busy hours.

What is WebRTC?

A web standard that lets apps do real-time voice, video, screen share, and data inside the browser. No installs. Works on laptops, tablets, and mobile. Adapts to changing networks.

Why it matters

- Join from anywhere with a link; reduces IT friction and wait time
- Low-latency two-way audio and video; adapts to loss and jitter
- Secure media (DTLS-SRTP) with options to keep traffic in-region depending on your infra
- Built-in support for screen share, recording, captions, and device access (camera, mic, scopes)
- Scales from 1:1 consults to group sessions via modern media servers

Where WebRTC actually runs, and what tends to break

Hospitals and large enterprises rely on steady, real-time communications. Here is where WebRTC shows up most often inside and outside healthcare:

- **Virtual care and remote monitoring:** High-volume visits, post-op follow-ups, home-based chronic care, and real-time monitoring with cameras and vitals (In other industries: field support, remote onboarding, customer walk-throughs, and CCTV)
- **Training:** Skills labs, live tutorials, external consultancy, and remote proctoring (In other industries: town halls, classroom labs, and vendor rollouts)
- **Clinical trials and research ops:** Omnichannel scheduling and referrals, decentralized visits, and remote site monitoring (In other industries: distributed project reviews and branch audits)
- **Imaging and diagnostics:** Teleradiology reads, remote ultrasound guidance, and lab result reviews (In other industries: media-heavy approvals, design reviews, and compliance sign-off)
- **Language and accessibility:** Medical interpreters, sign-language support, and real-time captions (In other industries: interpreter hotlines and captions for training and support)

Even the smallest defect can cause ripples to spread throughout your enterprise, leading to significant reputational, compliance, financial, and patient risk.

- **What tends to break in the real world:** Connects, then feels slow: While local area network (LAN) tools look green, end-to-end round trip time (RTT) is high on a longer backhauled path.
- **Heavy lip-sync and visible delay:** Media fell back to the transmission control protocol (TCP) or is relayed through traversal using relays around NAT (TURN) because the user datagram protocol (UDP) or ports are blocked, or the path hairpins through policy hubs.
- **Quality suddenly drops at peak:** Jitter and loss spike when imaging uploads, backups, or sync jobs overlap with calls.
- **Calls over 40 minutes drop near the same mark:** Idle timers or NAT quotas reclaim “quiet” flows.
- **Join fails on staff but not guest:** WebSocket upgrades or interactive connectivity establishment (ICE) checks get throttled by proxies on only one side.
- **One room or one SSID is always worse:** Last-meters issues like sticky clients, noisy radio frequency (RF) channels, or edge LAN quirks.
- **After a failover the app feels slower:** Region pinning drift increases distance and RTT.

Why this matters, and what comes next

These fingerprints usually trace back to a few realities. In the next sections, we unpack each one, show how to spot it quickly, and explain how to validate, observe, and prove fixes.

1) The path decides the experience

Traffic backhauling, hairpinning - VPN/MPLS/SD-WAN and cloud access security broker (CASB) - Zscaler, Prisma, Netskope

What it is: Branches send traffic to a central data center or policy hub instead of exiting locally. Security and compliance are centralized, but the path gets longer and queues appear.

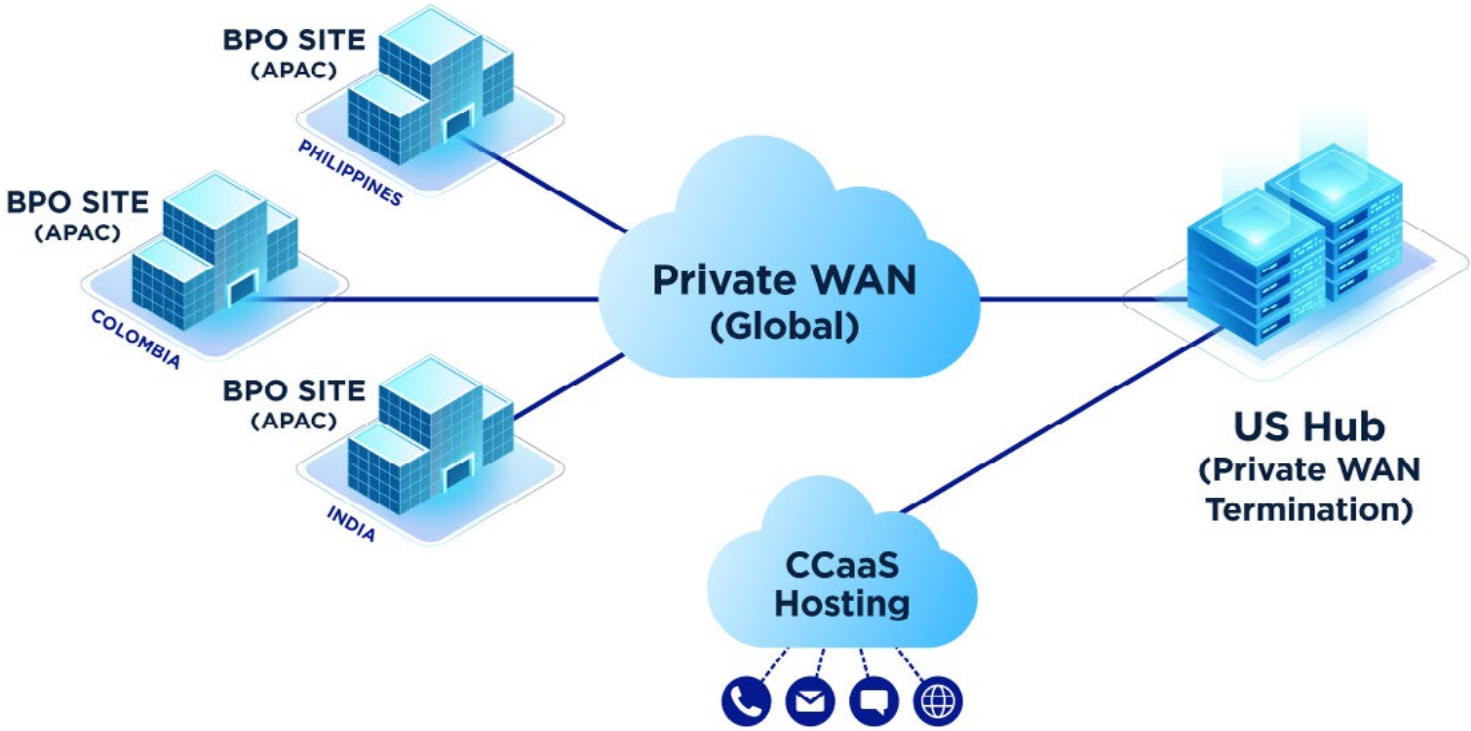
Why it matters for real-time communications (RTC): With secure service edge, the "hub" is often a cloud enforcement node. If your media is in AWS East and your SSE node sits in AWS West, packets detour to the SSE, then back to the media, adding latency and jitter. This is cloud hairpinning, and Microsoft's real-time media guidance explicitly warns against it. Live voice, video, and agent desktops are the first to feel it during peak hours.

SD-WAN reality: Changes in path weighting can shift calls onto thinner or busier circuits even when nothing has "failed." Everything is up, yet latency and jitter jump because traffic took the smaller pipe. Forward and return legs can also diverge. One direction looks clean; the other adds hops and trouble. You'll see this as jitter bursts and fresh ICE activity during otherwise normal operations.

Typical signs of traffic backhauling:

- LAN tools look green yet call RTT is high.
- Jitter spikes line up with peak times on the backhaul.
- Two similar sites diverge after a policy push or failover because region pinning changed the exit path.
- One-way quality complaints when forward and return legs use different SD-WAN routes.
- Differentiated services code point (DSCP) marking gets stripped in the backhaul, so media competes with imaging and backups at peak.

Compliance consideration (data residency/region): Healthcare and finance often require traffic to stay in approved regions. After a failover or policy change, routing can drift. Latency rises and compliance risk appears. Verify the region sessions actually use, not only the one intended.



Takeaway

Backhaul sets up the road while NAT sets up the gates. Together they define how far you can scale before long calls and busy-hour traffic drops.

CYARA

2) The gates decide the headcount

NAT, CGNAT, and double NAT

NAT cheat sheet

- **Network address translation (NAT):** The branch or home edge rewrites private addresses to a public one and keeps a per-flow state.
- **Carrier-grade NAT (CGNAT):** The internet service provider (ISP) does another layer of translation, so many customers share one public IP with per-IP and per-port quotas.
- **Double NAT:** Both layers at once.
- **Symmetric NAT:** Mappings change per destination, which makes direct peer paths rare.

Why this matters

- **WebRTC prefers UDP end-to-end:** Extra translation pushes more media to TURN relays instead of direct. That adds delay and cost.
- **Quotas and tables:** CGNAT and firewalls enforce per-IP and per-port limits. At peak, new calls flap or fail.
- **Idle timers:** Stateful devices close “quiet” flows on fixed intervals, so long calls die at repeatable minute marks and reconnect.
- **Hot spots:** When traffic is backhauled, a few NAT devices carry many sites. Tables fill, mappings get recycled, and stability drops.
- **Firewalls are built to conserve energy:** They aggressively age out UDP sessions that look idle (mute).

Typical signs

- **Heavy lip-sync and noticeable delay even though the LAN looks fine.** Media is relayed via TURN instead of a direct path.
- **Calls over 40 minutes drop near the 45-minute mark.** Long calls fail at repeatable marks because stateful devices reclaim mappings. Common patterns are 15, 30, 45, or 60 minutes. Keepalives help, but tight timers or CGNAT quotas will still cap concurrency and session length under load.
- **People cannot join at peak but can off-peak.** Port or connection quotas are being hit behind CGNAT or a busy firewall.
- **Some branches on a particular ISP struggle, while home or guest Wi-Fi is fine.** CGNAT or strict NAT at those sites.
- **Intermittent “can’t connect” that clears after a few minutes.** NAT tables are recycling mappings under load.
- **Relayed percent is unusually high; ICE restarts spike; staff SSID fails, while guest works.** Symmetric NAT, blocked UDP, or policy pushing calls onto relays.

Takeaway

The road sets distance and delay. The gates set concurrency and session length. If you do not test under sustained, concurrent load, these limits only show up in production.

3) Bandwidth under real load (not just a speed test)

Why a typical speed test misses reality

Average bandwidth is a deceptive metric: in healthcare, imaging syncing creates micro-bursts milliseconds of intense traffic that overflow switch buffers. This causes tail-drop packet loss that standard monitoring tools average out and miss, but WebRTC feels instantly as robotic voice.

Typical signs of bandwidth issues:

- Fine at 50–80 users; quality falls around 90–100
- Loss ramps after 5–15 minutes as queues fill
- Screen share blurs or stalls when imaging or file uploads start
- Long consultations degrade over time as devices warm up, buffers grow, and NAT quotas or timers begin to bite

A short test cannot reveal where backhauling adds delay or where NAT and CGNAT impose quotas and timers. You need a busy-hour soak that runs your real media mix, from your sites, along the actual policy path, long enough to find the breakpoints.

Speed test vs. reality

- **Wrong path:** Most public speed tests hit nearby, well-peered endpoints. Your real-time traffic often goes through VPN or SD-WAN backhaul, TURN, and policy hubs. Different path, different result.
- **Wrong protocol and ports:** Most speed tests are short TCP bursts to a CDN. WebRTC is sustained UDP with control traffic and renegotiations. If the test never uses your media ports or TURN, it will not expose those limits.
- **Too short, no concurrency:** A 10–20 second test never fills queues or heats devices. Real sessions run 30–90 minutes, with dozens or hundreds in parallel. That is when NAT quotas, idle timers, and backhaul queues show up.
- **Bypasses the pressure points:** ISP policy routing and QoS markings can put speed tests on a cleaner class than your media. The test looks great, but production does not.

CYARA

4) Wi-Fi and edge LAN reality

What it is:

Carts, tablets, and bedside stations get parked and used anywhere on campus. The same device can behave very differently by room or floor because the radio environment and the edge LAN are not uniform. You see the same thing in contact centers and large offices with many meeting rooms.

Typical signs:

- One room experiences issues regularly, while another room is consistently better.
- The same cart that performs well on Floor 2, has sudden issues when used on Floor 5.
- Desktops on the same VLAN look fine, but the cart in Room 512 doesn't.
- "It works on guests, not on staff" when SSID/AP choice changes.

A short test cannot reveal where backhauling adds delay or where NAT and CGNAT impose quotas and timers. You need a busy-hour soak that runs your real media mix, from your sites, along the actual policy path, long enough to find the breakpoints.

Why it matters for RTC

Quality changes between sessions: Different Wi-Fi received signal strength indicator(RSSI)/signal-to-noise ratio (SNR), 2.4 vs 5/6 GHz, noisy channels, and busy zones (wards, lobby, radiology corridor) can all affect performance. Hospital Wi-Fi coverage and usage density aren't uniform, so "yesterday was fine" doesn't guarantee today's quality.

Why these show up in RTC

Sticky clients: Devices cling to a distant access point instead of roaming to a closer one.

- Impact on RTC: Rising delay and bursty loss and you'll see lip-sync drift and micro-mutes.

RSSI and SNR: RSSI is the signal level the client hears; SNR is how loud that signal is compared to the background noise.

- Impact on RTC: Low signal or poor SNR means more retransmissions on the air, which turns into jitter spikes and short freezes in voice and video.

Band and RF channel: 2.4 GHz is crowded with few clean channels; 5 GHz offers more room but is still shared.

- Impact on RTC: A busy or overlapping RF channel adds airtime waits and retries, so media frames arrive late or bunched.

DFS channel moves: Some 5 GHz channels must vacate if the AP detects radar, etc. The AP shifts clients to a new channel automatically.

- Impact on RTC: a brief hitch or freeze right when the move happens, then recovery. This explains "good all day, one sudden wobble" reports.

Channel utilization and retries: Utilization is how busy the RF is; retries are frames sent again because the first copy was corrupted or collided.

- Impact on RTC: High utilization or many retries produce exactly the "fine one minute, choppy the next" behavior.

Edge LAN health: Aging switches, power issues on APs, access control lists (ACLs) that treat staff and guests differently, or DHCP scope problems.

- Impact on RTC: It looks like "Wi-Fi is bad," but the bottleneck is a switch port, a policy on the staff SSID, or an address assignment hiccup.

Identify patterns with Cyara's solutions:

- Use **watchRTC** custom keys to tag sessions with device and room IDs. If your client allows it, also record the AP identifier and basic radio health like RSSI and SNR.
- Place **probeRTC** monitoring nodes near known hot spots to keep a pulse by shift and busy hour.
- Run **qualityRTC** pre-call checks at the point of use, so weak signal or high round-trip time is caught before a consult starts.

Why this is not only a hospital problem:

Swap "ward" for "contact center pod" or "town hall room." The last meters behave the same in large offices and BPO floors: mixed AP vendors, uneven AP placement, and per-SSID rules that differ for staff vs. guests.

If rooms and carts vary even when the path is clean, be sure to check the hidden bucket.

5) The hidden network bucket

The following issues often travel together. And while they each may appear minor, they can cause a call that passed a lab test to fail on a clinic floor.

Transport

What is happening in this layer:

- **UDP blocked → media over TCP/TLS:** WebRTC is designed to carry voice and video over UDP. Some networks block UDP or media ports, so browsers fall back to TCP or TLS over 443. TCP guarantees delivery in order, which is great for files but terrible for live audio and video. Late packets still get delivered and pile up in buffers. The call starts to lag, lip-sync drifts, and screen share "tears" because frames arrive in clumps.
- **TLS inspection or proxying in the path:** Many enterprises use middleboxes that intercept TLS to scan traffic. WebRTC media typically uses DTLS-SRTP over UDP, which those boxes cannot inspect. Results vary: some devices block DTLS outright, some rewrite certificates or ciphers, and some interfere with WebSocket upgrades used by signaling. You get odd "handshake" or "upgrade" errors even though ordinary browsing works.
- **Proxy PAC rules for signaling:** Auto-config scripts can steer WebSocket signaling through SSL gateways even when media goes direct. So joins hang or renegotiations flap only on the corporate LAN.

Why RTC feels bad when transport is wrong:

- **Head-of-line blocking with TCP:** One lost packet holds up all later packets, so interactive media accumulates delay.
- **Bufferbloat:** TCP keeps the pipe full. When queues inflate, you hear long talk-over and see frozen frames followed by bursts.
- **Relay side-effects:** If UDP is blocked, sessions often route through TURN over TLS. That adds distance and jitter even when the nominal bitrate is modest.
- **Handshake flaps:** TLS interception can break DTLS keying for media or block WS upgrades for signaling, so calls connect in the lab and fail on the floor.
- PAC-steered signaling stalls ICE or drops keepalives while general web traffic looks fine.

Fingerprints

- **Relayed percent is high** where it shouldn't be; flows show only 443/TCP for media
- **ICE completes but quality is spiky** at normal bitrates (classic TCP fallback)
- **Signaling fails in production, not lab** (blocked/throttled WS upgrades)
- **Quality drops right after a policy change;** transport flips from UDP to TCP, relayed percent jumps across multiple sites

Packet handling

What is happening:

- **Shrunk maximum transmission unit (MTU) and fragmentation from tunnels and overlays:** IPsec, GRE, or VXLAN add headers, so the effective MTU gets smaller. If path MTU discovery is blocked, oversized packets are fragmented or dropped. Flows “connect,” then wobble because big media packets never make it end to end.
- **Idle timeouts on stateful devices:** Firewalls, proxies, and NATs expire “quiet” flows on fixed timers. Long consults hit those timers and the session gets torn down, then immediately reconnects.

Why RTC quality feels poor:

- **Video and screen share break first:** Video and screen share packets are larger, so MTU problems hit them before audio. You see stalled or frozen frames while audio limps on.
- **Fragmentation creates jitter and reordering:** Extra splits and reassembles mean bursts of late or out-of-order packets, so calls “connect but never settle.”
- **Timer cuts are surgical:** Sessions die at exact minute marks that match device policies, not user behavior.

Fingerprints:

- Bursts of loss and out-of-order with no CPU or bandwidth spike on endpoints.
- Drops at 30, 60, 90, or 120 minutes like clockwork.

Routing and name resolution

What is happening:

- **Split-horizon DNS:** Internal resolvers give different answers by network. A clinic may resolve your media hostnames to a far region or to names that only exist inside staff LANs.
- **Overlapping private IP space across VPNs:** Two sites use the same RFC1918 ranges, routes collide, and packets take the wrong tunnel or bounce between them.
- **IPv4 and IPv6 asymmetry:** Dual stack often means HTTPS signaling prefers IPv6 while media prefers IPv4. NAT64 and DNS64 on v6-only segments can also rewrite things in flight. Some CPaaS stacks still negotiate media only on IPv4, so families split.

Why RTC quality feels poor:

- **Wrong region, higher RTT:** Calls connect but feel sluggish because DNS steered media to a distant point of presence (PoP).
- **Blackholes and hairpins:** Overlapping space creates ambiguous routes. Meanwhile, media flows take a path that cannot return, or loop through the wrong hub.
- **Family mismatch:** Signaling succeeds over IPv6, but media candidates over IPv4 never pair cleanly. ICE takes longer, falls back to relays, or fails for certain OS and network combos.
- **Resolver quirks:** Staff DNS blocks or times out on some hostnames, so joins hang while guest works.

Fingerprints:

- One clinic is always worse, even with new devices. Traces show media egressing in a farther region.
- “Host not found” or slow DNS only on the staff SSID. Guest is clean.
- Calls succeed after you switch to a public resolver.
- After a failover, RTT jumps and never returns to baseline.
- ICE takes longer and lands on relayed paths at one site, direct paths elsewhere.

VDI and thin clients (Citrix or VMware)

- **Setup:** Many clinicians launch telehealth inside Citrix or VMware Horizon to keep data central.
- **Problem:** Without browser content redirection (BCR), the video is rendered on the VDI server, re-encoded, and sent to the user as a "video of a video."
- **Cost:** Server CPU burns, hardware acceleration is lost, lip-sync drifts, and frame rate falls, even on a clean network.

Local firewall and endpoint security

- **What happens:** Endpoint security and local firewalls insert themselves into the browser's network stack. SSL inspection and content scanning add per-packet overhead. CPU spikes create "pseudo packet loss" where the network path is clean but frames are late or dropped. Some agents also push media to TCP/TLS by blocking UDP.
- **Fingerprints:** TURN over TCP even when the site path is open, bursty jitter with no WAN congestion, high browser "network service" or AV process CPU during calls, good stats on quick lab checks but choppy production sessions.

What can you do?

These common WebRTC issues do not have to plague your healthcare organization. Instead, you can leverage Cyara's solutions to identify issues, accelerate troubleshooting with continuous monitoring, and optimize your infrastructure for success.

1 **Prove capacity under real load**
Run the Cyara testRTC Network Saturation tool from your sites, through your backhaul, for the durations and concurrency that matter. Anchor it with Cyara Virtual Assistant (CVA) on-prem and Cyara Cruncher to push real mixed media. Add your usual background jobs.

Outcome: You will see the true breakpoints, including queue build-up, NAT quotas, timer cuts, and the first quality dip at busy hour.

2 **Qualify each site for readiness**
Use Cyara qualityRTC as a pre-join or a site turn-up step, so the test rides your actual path: VPN or SD-WAN, proxies, region pinning, NAT, and DNS. It surfaces UDP and TURN reachability, port blocks, and MTU hints.

Outcome: You'll receive a clear yes or no on "will calls set up and be stable here right now."

3 **Keep a pulse between changes**
Place probeRTC monitoring at critical clinics, BPO floors, and edge links. When SD-WAN weights shift or a new policy lands, you see it the same day instead of a week later.

Outcome: Drift cannot hide.

4 **See production truth and drill down**
Enable Cyara watchRTC in the app and tag sessions with the keys you already know: device or cart ID, room, OS, SSID, site. Patterns and repeat offenders jump out, and you can verify fixes and catch regressions fast.

Outcome: Make more informed decisions based on real-live performance, not lab guesses.

Conclusion

At 10:00 a.m. on Monday, when the clinic is full and the network is red-lining, your monitoring dashboard shouldn't be a mystery novel. That moment is not for guessing; it's for patient care.

We've outlined many risks that can arise when RTC performance takes a turn for the worse: longer paths from backhaul, tight NAT gates, peak-hour queues, and last-meter variance. The way to stay ahead is simple and repeatable: test the real path at scale, qualify each site, keep a pulse, then watch production with enough context to separate room, device, and network issues.

Simply, the path matters. The gates matter. Peak hour tells the truth. A few hidden quirks can ruin clean lab graphs. The fix is proactive, not reactive. Cyara helps you prepare for these realities and mitigate the risks to healthcare organizations and enterprises across all industries.

Beyond just testing your network, Cyara helps you validate journeys before they go live, stress-test the gates, and watch the reality of every cart and room in real-time. When you make proactive RTC testing and monitoring a priority, you cut escalations, speed rollouts, keep traffic in the right region, and assure voice and video quality for when it matters most.

Now's the time to make busy hours uneventful for your IT and support teams.

Contact us and see how you'll benefit from the Cyara Agentic Platform. Learn more today.

Learn more