Conversational Care: Designing Responsible Human-AI Care Coordination

Marco Da Re
Dyson School of Design Engineering
Imperial College London
London, United Kingdom
m.dare@imperial.ac.uk

Pablo A. Fonseca
Dyson School of Design Engineering
Imperial College London
London, United Kingdom
p.fonsecaarroyo@imperial.ac.uk

Minseo Cho
Dyson School of Design Engineering
Imperial College London
London, United Kingdom
m.cho24@imperial.ac.uk

Javier Sanchez
CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia,
Lima, Peru
javier.sanchez.c@upch.pe

Alina Irina Serban
UK Dementia Research Institute, Care Research and Technology Centre
Imperial College London
London, United Kingdom
alina-.serban18@imperial.ac.uk

Francisco Diez-Canseco
CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia,
Lima, Peru
fdiezcanseco@upch.pe

Maria Sofia Cuba-Fuentes
Center for Research in Primary Health Care
Universidad Peruana Cayetano Heredia
Lima, Peru
maria.cuba@upch.pe

Ravi Vaidyanathan
Department of Mechanical Engineering
Imperial College London
London, UK
r.vaidyanathan@imperial.ac.uk

Rafael A. Calvo
Dyson School of Design Engineering
Imperial College London
London, UK
r.calvo@imperial.ac.uk

Healthcare systems worldwide face demand they cannot meet, and conversational artificial intelligence (CAI) is one of the pathways to scale provisions. But doing so responsibly requires addressing safety, fairness, effectiveness, accessibility, and several other criteria at once. Few projects have the resources to build a system that meets all of them from scratch. We followed a 2-year Research through Design process to iteratively build *Conversational Care*: a customisable platform that integrates AI agents into care-team workflows, automating routine tasks and escalating to humans when oversight is needed. The platform embeds nine responsible-AI criteria drawn from the literature. It has been designed with, and to support, community health workers, nurses, and families of people living with dementia in Peru and the UK; health coaches and link workers in London; with early exploration of care home monitoring in Singapore. Working across such varied scenarios surfaced the design trade-offs and recurring patterns that emerge when CAI is integrated into real care systems – contributions that other teams can adapt to their own contexts.

Conversational AI, Community Health, Research Through Design, Artificial Intelligence, Application Framework, AI Agents, Chatbots, Care Coordination

1. INTRODUCTION

Across the globe, demand for health services far outstrips available supply. Technology, including AI in its many forms, is therefore a logical pathway for scaling provisions.

The design challenges of doing this responsibly are particularly acute in the implementation of agentic Conversational Artificial Intelligence (CAI), including chatbots, voice agents, and other systems that use natural language to act on behalf of, or alongside humans.

CAI raises issues that are different from those in other machine-learning applications for health (Wiens et al. 2019), where the focus is mainly on data quality and model performance. CAI systems are social agents which reshape the relationships between healthcare providers, patients and their unpaid carers. For example, what happens when ChatGPT gives advice to patients that is different to that from their doctors, or family members? Or when clinical teams try to delegate tasks to CAI? CAI design therefore carries serious ethical consequences that must be considered from the start (Car et al. 2020; Sundareswaran and Sarkar 2020).

A review of 101 papers by (Meadi et al. 2025) lists the criteria most discussed for responsible CAI: safety, explicability and transparency, responsibility and accountability, empathy and humanness, anthropomorphisation and deception, fairness and justice, and autonomy. Less commonly discussed are issues specific to CAI's social nature. For example, the impact on the work and motivation of paid and unpaid healthcare providers whose tasks the system partly absorbs.

Organisations are trying to understand how to scale healthcare services with AI while respecting ethical criteria, meeting stakeholders needs and sustaining the motivation of human stakeholders. Individual projects rarely have the funding or time to elicit stakeholder needs, design the interactions, build a software platform, and evaluate against every relevant criterion.

One alternative is to embody the experience and best practices learned across multiple projects in an *application framework* (Fayad and Schmidt 1997): reusable design and software components that can be adapted to different contexts.

This paper presents *Conversational Care*, a working software platform built as one such application framework. We use it to show how nine responsible-AI criteria become concrete, reusable design decisions, and what recurs when CAI is built into real care settings.

The platform integrates AI agents into care-team workflows, aiming to improve human-AI coordination in healthcare contexts. It was developed iteratively and deployed across the UK and Peru, with stakeholder research extending to Singapore. Throughout the paper we use "platform" for the deployed system itself and "application framework" for the software-engineering approach behind it.

To understand the stakeholder needs common to such a domain and the concrete ways to address them, we use a Research through Design (RtD) approach, which treats design practice itself as a way of generating knowledge. Rather than only evaluating the finished AI tool. RtD lets us learn from the iterations, tensions, prototypes, and stakeholder interactions that emerge during design.

Section 2 sets out nine responsible-CAI criteria identified in the literature. Section 3 describes the RtD process for satisfying these criteria. Section 4 presents the design decisions that resulted, organised by criterion, in ways that could be applied to other projects. Section 5 discusses what cross-deployment work revealed, and Section 6 concludes contextualising the implications.

2. RESPONSIBLE CAI IN HEALTHCARE

Wellbeing-led design integrates ethical and wellbeing criteria into design practice (Peters et al. 2020). For this paper we adopt the criteria proposed by Zeng et al. (2026) for responsible AI in primary care.

This taxonomy is particularly relevant here because the deployments described in this paper sit mostly in primary care, rest on therapeutic relationships that are hard to capture in standard metrics and often run in low-resource settings.

We extend Zeng et al.'s set with another criterion - **conviviality** - drawn from the wellbeing-led design tradition (Peters et al. 2026). It addresses the autonomy-supportive quality of human-AI relationships, which the other criteria do not cover directly. Together, these form the nine criteria used throughout this paper.

(i) **Safety**: operating without unacceptable risk of harm to patients, clinicians, or the health system. This spans relational issues such as reciprocity (Wieland 2023) and technical ones such as model-accuracy degradation in dynamic clinical environments.

(ii) **Fairness**: performing equitably across population groups, including at intersections of multiple marginalised identities. Sadek et al. (2023) review best practices for designing CAI equitably, most using participatory approaches.

(iii) **Effectiveness**: improving patient outcomes, clinical processes, or system performance compared to standard care. Reviews before (Milne-Ives et al. 2020) and after (Qin et al. 2025) the launch of ChatGPT show consistently weak outcome evidence, partly because robust evaluation requires complex randomised control trials (RCTs).

(iv) **Generalisability**: performing across diverse populations, settings, and time periods beyond the original development environment. AI benefits have been consistently overstated, even before large language models (LLMs). Nagendran et al. (2020) found few RCTs exist, most at high risk of bias or limited real-world relevance.

(v) **Usability and Accessibility**: all clinicians, staff, and patients can effectively, efficiently, and satisfactorily use the AI system. Project success often depends on the management interface used by care-teams overseeing many patients. Yet there is scant evidence on how these decision support systems should be designed (Wang et al. 2023).

(vi) **Human-centredness**: paying explicit attention to the preferences, values, goals, and social contexts of their intended end users. Sadek et al. (2023) synthesise value-sensitive design approaches for CAI and offer co-design recommendations.

(vii) **Coordination**: facilitating care across settings, providers, services, and time. Recent work on human-AI coordination in healthcare is still within single contexts (Sadek et al. 2023). Solano-Kamaiko et al. (2026) examine how CAI might support coordination among the multiple carers – paid staff, clinicians, and family – involved in home-based care. Some propose these systems to become "thought partners" (Collins et al. 2024), but these models say

little about coordination when multiple agents are doing the "thinking".

(viii) **Conviviality**: supporting autonomy and the collaborative relationships that reduce friction and strengthen the human connections that make care effective. The term is from Illich and Lang (1973): convivial tools augment human agency and relationships rather than replacing them, unlike "industrial" tools that create dependency and narrow autonomy. Molden (2024) discusses the limits of automation in relational care disciplines; Peters, Espinoza, Da Re, et al. (2026) examine conviviality for healthcare in low- and middle-income countries. Evidence to date suggests CAI tends to absorb specific tasks rather than entire roles (Bonney et al. 2024), and its effect depends on which tasks are redistributed (Eloundou et al. 2024).

(ix) **Continuity**: supporting continuous, longitudinal care relationships and information continuity over time, across life stages, and through changing health states. Its evaluation is itself a gap in the healthcare CAI literature.

While the literature has identified multiple challenges across these categories, those designing and building healthcare systems need best practices to address them. The criteria are difficult to study independently, as changing one can affect another, though project scope often leaves no other choice.

Developing conversational AI for healthcare is therefore a "wicked problem" (Rittel and Webber 1973): multiple stakeholders (patients, clinicians, unpaid carers, managers, regulators, vendors) with conflicting values (safety vs. access, efficiency vs. relational care, standardisation vs. individualisation), under high uncertainty (emergent LLM capabilities, evolving regulation, shifting workflows). Rather than addressing the nine criteria abstractly, we use Research through Design to surface them empirically. Through the process, we see what breaks, what works, and what unexpected uses emerge once a concrete system is introduced.

3. RESEARCH THROUGH DESIGN

We address the nine criteria by adopting Research through Design practice (Zimmerman et al. 2007). RtD treats the act of designing – framing, ideating, prototyping, deploying, and evaluating – as a mode for knowledge production in its own right. Design artefacts (journeys, prototypes, scenarios) act as probes into complex situations where problem and solution co-evolve. This let us translate the nine criteria into concrete artefacts while surfacing tensions, opportunities, and unintended consequences along the way.

3.1 Process overview

From June 2024 to May 2026, we ran a participatory RtD process across sites in Peru, and the UK, structured in three phases: (1) stakeholder research and framing, (2) iterative design, and (3) PPI-led evaluation and refinement.

Scenarios spanned dementia care, social prescribing, health coaching, and care home monitoring. Different as they are, they share key features: all involve care recipients, formal and/or informal carers, and a monitoring team; all require structured conversations (information gathering, scheduling, care transfer, information provision) across modalities (WhatsApp, SMS, calls); all sit below clinical urgency; and all surface the same monitor concerns: lack of time and need for motivation.

The work began as two separate projects: *Impact Salud* in Peru and *UK DRI Minder* in the UK. In mid-2025 their parallel technical workstreams converged into a single shared platform (Section 3.3). Later deployments for social prescribing in the UK (Cho et al. 2026) configured that platform rather than building anew. In Singapore, university partners ran an early-stage collaboration on sensor-integrated conversational agents for care home monitoring - supporting staff who look after older adults ageing in place. No deployment has run there yet, and we give no further detail here; the collaboration will be documented separately (https://snowwon.github.io/nus_icl_hackathon/).

The process engaged 145 stakeholders across 14 roles (Table 1) - care navigators, community health workers and monitoring teams in Peru and the UK; clinicians, carers, and people living with dementia in the UK; and clinical researchers across Peru and Singapore.

Table 1: *People involved in the RtD Process*

Role	Location	#	Contribution
Care Navigators Supervisor	*Peru (Lima)*	*1*	*Defined operational requirements and team workflows for the Peru deployment; co-designed the monitoring platform and its agents.*
Care Navigators	*Peru (Lima)*	*2*	*Shaped interface choices for the monitoring platform through daily use and iterative feedback.*
Community Health Workers and Field	*Peru (Lima, Iquitos, Huancay*	*4*	*Brought culturally sensitive perspectives across four contrasting Peruvian sites,*

Workers Supervisors	*o, Tumbes)*		*informing how to approach community members locally.*
Community Health Workers and Field Workers	*Peru (Lima, Iquitos, Huancay o, Tumbes)*	*40*	*Informed operational logistics and admin needs of frontline work; in design sprints, reflected on whether automation eroded or sustained their sense of purpose.*
Carer and Dementia Advisor	*Peru (Lima)*	*1*	*Provided personal and contextual expertise on dementia care in Peru.*
Monitoring Team for Dementia Care	*UK (Hammer smith & Fulham, Surrey)*	*8*	*Informed case management and longitudinal monitoring requirements in the early UK pilot.*
Clinical Monitoring Team for Dementia Care	*UK (Hammer smith & Fulham, Surrey)*	*4*	*Reviewed escalation pathways and clinical oversight needs.*
Clinicians (GPs Occupational Therapy, UK DRI)	*UK (London)*	*10*	*Contributed clinical workflow knowledge and informed product development for the UK healthcare context.*
Younger Adults, Social Prescribing	*UK (London)*	*12*	*Informed the design of CAI for social prescribing contexts.*
Link workers West London Trust	*UK (London)*	*8*	*Informed the design of CAI for social prescribing contexts and future implementation in care practices.*
Clinical Researchers	*UK (London), Peru*	*15*	*Informed evaluation design across sites.*
People living with Dementia	*UK (London)*	*20*	*Shaped accessibility choices through PPI activities (voice and text modalities, slower pacing, accent options) and identified where automation risked eroding human connection.*
Carers	*UK (London), Peru*	*15*	*Shared lived experience of caregiving burden, informing which informational tasks could be automated.*
Design Engineers	*UK (London)*	*5*	*Team of designers and engineers who built the technical infrastructure and coordinated the development of Conversational Care.*

3.2 Phase 1: Stakeholder research and problem framing (June – December 2024)

The first LLM-powered CAI system built by this team drew on Espinoza et al.'s system, *Ana* (Espinoza et al. 2025), and a text-based conversational agent for asthma support (Cook et al. 2024). These surfaced common requirements: participant and agent management, and the human-to-human coordination that hybrid (human–AI) systems require.

From these common requirements, we set out to design one system that could be adapted to various contexts. *Conversational Care* began as a management system for virtual assistants, bringing together human-led phone calls and AI-assisted communication on a single platform. Phase 1 scoped it around the *Impact Salud* project in Peru (Lazo-Porras et al. 2025), the first of several studies the platform now supports.

Fieldwork combined ethnography (Crabtree et al. 2012) and contextual inquiry (Viitanen 2011) across four Peruvian sites - Lima, Iquitos, Huancayo, and Tumbes (Porras et al. 2025) - run in parallel with the deployment of an mHealth tool for community health workers (Da Re et al. 2025; Espinoza et al. 2024). Because the same stakeholders participated in both the fieldwork and the mHealth deployment, the platform's early requirements came from people whose practice we already knew.

In August 2024, a design sprint (Knapp et al. 2016) with the *Impact Salud* team mapped the Care Team Navigator (CTN) journey across two stages (recruitment/enrolment and day-to-day care) identified five roles (Coordinator, Nurse Supervisor, CTN, Field Worker/CHW, Family), and located the related system touchpoints in the care flowchart (Figure 4, Appendix). Features were sketched collectively, then prioritised by clinical and tech stakeholders against expected impact and development effort (Figure 5, Appendix). The outputs fed into the first design iterations (Figure 6, Appendix).

In parallel, the UK pilot focused on dementia care within the *UK DRI Minder* project (Serban et al. 2026). Patient and Public Involvement (PPI) sessions with designers, monitoring team members, occupational therapists, and researchers mapped the team's daily tasks onto a Time vs. Value matrix. The matrix surfaced tasks that were time-consuming

for the team but low value for people living with dementia (PLwD) and their study partners. The aim was to free the team to spend more time on higher-value work.

Phase 1 produced sprint outputs (CTN journey map, five role definitions, feature backlog, prioritisation matrix), ethnographic field notes from the four Peruvian sites, and interview transcripts.

3.3 Phase 2: Iterative design and implementation (September 2024 – August 2025)

Phase 2 ran two parallel technical workstreams: the *Impact Salud* deployment in Peru, and a proof-of-concept prototype (Minder Chat) for the *UK DRI Minder* project.

For *Impact Salud*, the technical team developed a first prototype, meeting weekly for 4–6 hours with intervention coordinators. Implementation in Peru proceeded in two stages. In February 2025, a higher-fidelity prototype using Django (a Python web framework) for the app and Airtable (a cloud database) for custom forms was presented to intervention coordinators for validation and iteration. In April 2025, once the forms were validated, key functionality was integrated: phone calls via Twilio (a cloud communications API) and a custom content management system (CMS) built directly into the app. A final version including the chatbot was produced in November 2025 for the start of the *Impact Salud* intervention (Section 3.4).

For the *UK Minder Chat* prototype, work ran from August to December 2024. Built in Django, it used Airtable for data storage, Twilio for both human-led and automated phone communication, and a retrieval-augmented generation (RAG) conversational agent with a classifier node built on Open AI's GPT models to route incoming messages to the relevant branch in the conversation.

Both workstreams converged on the same core stack - Django, Airtable, and Twilio - diverging only at the functionality layer. This shared foundation, consolidated in mid-2025, is what became *Conversational Care*.

3.4 Phase 3: Piloting, Iterations and Reflection (September 2025 – May 2026)

The first real-world deployment ran in Peru with *Impact Salud*, pairing phone calls led by care team navigators with a WhatsApp RAG assistant for out-of-hours support, reaching 64 families of people living with dementia.

The platform was then extended to UK studies on dementia care, reaching 90 users who tried the platform or its agents at least once. A separate UK study on social prescribing involved 14 young people and 13 link workers (Cho 2025). Across all deployments, around 180 people have used the platform to date (July 2026).

Alongside these deployments, the team monitored performance, iterated on functionality, and adapted the platform for new use cases.

At the **strategic** level, we aligned the core team and shaped the product vision via a Foundation Sprint (Knapp 2025) in September; surfaced new use cases via partner interviews across Peru, the UK, and Singapore; and learned from existing community-care tools via a literature review (Cho et al. 2025).

At the **platform** level, we iteratively improved usability via cognitive walkthroughs (Nielsen 1994) with active users and heuristic evaluation (Nielsen and Molich 1990), addressing navigation, missing features, and safety/fairness concerns.

At the **agent** level, we ideated and prototyped virtual assistants via co-design sessions with people living with dementia and testing with PLwD and carers (together and separately), covering scheduling, preparing PLwD to leave the house, 24/7 dementia information, and signposting to local services. We also identified admin and repetitive tasks that could be lifted without substituting for human care, via prioritisation work with monitoring teams. This conversation is ongoing, since hands-on experience with the tools gives care teams the clearest sense of what is worth automating.

3.5 Analysis

Analysis asked whether, and how, the evolving platform met the nine criteria. It was iterative and reflective, consistent with RtD, where rigour lies in design rationale and documentation rather than retrospective coding (Zimmerman et al. 2007).

The analysis drew on the materials the process produced: design artefacts, workshop and interview transcripts (discussed in detail by Peters et al. 2026), ethnographic field notes from the four Peruvian sites, prototype versions with decision logs, and platform telemetry with structured pilot records. Together, these link each design choice to the evidence that shaped it.

The criteria acted as the recurring analytic lens. Each was operationalised as a question we asked of design decisions (e.g., Safety: does this change the risk of an unsafe escalation? Fairness: does it work across the language, literacy, and modality preferences observed in our sites? Conviviality: does it preserve the relational quality of the work that human staff value?). At review meetings, we walked through recent decisions, prototypes, and pilot observations against the criteria, flagging which were addressed, which surfaced tensions, and which remained open. The criteria were a continuing

prompt during design rather than a post-hoc checklist.

This review ran at two levels. **Within each project**, technical and design teams met regularly to review what had been tested, what had failed, and how stakeholder input was reshaping the framing; decisions and their evidence base were recorded in project documentation. **Across projects**, the core team compared emerging patterns: which design choices recurred, where contexts diverged, and which problems were shared. The cross-site view shaped what entered the platform (Section 4) and what stayed deployment-specific. The unresolved tensions are taken up in Section 5.

3.6 Ethics

This paper draws on results from three separately approved studies: *Impact Salud* [ICREC 6647445] and *UK DRI Minder* [ICREC 7610107] and *Reflective Companion (RECO)* [ICREC 7558516]. The remaining work with researchers and clinicians took the form of collaborative research activities. Informed consent was obtained for all recorded activities. We report how these conversations informed design decisions.

Recruitment ran through established community networks. Community partners experienced with PLwD co-facilitated the workshops, handled logistics, and supported capacity-to-consent assessment. PLwD with capacity to consent signed consent directly; where capacity was limited, their carer signed on their behalf. Participants received hourly compensation in line with each project's guidelines.

Generative AI tools (Claude and an institutional writing assistant) were used for grammar and copy-editing only.

4. SYSTEM DESIGN OUTCOMES

Conversational Care is a deployed, white-label agent management platform for human and AI care coordination, applicable to various care settings. “White-label” means other teams can adopt, rebrand, and configure it for their own deployments. In software-engineering terms, the platform is built as an “application framework” (Fayad and Schmidt 1997): it embeds best practices for managing human participants (patients, clinical teams), managing or linking to conversational agents, and orchestrating both in healthcare contexts. Across the three projects where it has been used to date, it enables care teams (monitors, care navigators, link workers, or clinicians) to scale the number of people they support by automating routine and administrative tasks, escalating to humans only when oversight is needed.

The platform addresses problems that recur across care contexts, while allowing customisation for each setting. Each deployment combines and configures features from a shared set: which channels are active (WhatsApp, SMS, voice, web), which agents are enabled, and the thresholds at which the system escalates to human review. Its reusable components reduce development costs and embed best practices.

The platform (Figure 1, Appendix) brings together the tools care teams use day-to-day: placing calls, triggering alerts, scheduling appointments, collecting data and notes, reviewing conversations and alerts in real time, and accessing patient history. From the platform, various AI automations can be triggered. Automations vary in scope according to the care team’s needs. Currently developed agents (Figure 2, Appendix) collect data, coordinate follow-ups, send reminders, and provide 24/7 information on local services or specific health topics, escalating to the team through a classifier when appropriate.

In the following sections we describe the design decisions the RtD process produced across the nine responsible-AI criteria, tracing the system’s features back to the design work that produced it. A complete description of the platform's interface features is beyond the scope of this paper; these are maintained on the platform website (https://conversational-care.ai) and in future papers on specific deployments.

4.1 Safety

Escalation pathways and clinical oversight were the safety needs most consistently raised by monitoring teams and clinicians in Phase 1 (Table 1). In response, we implemented a **multilevel safety approach** that combines proactive testing, automated safeguards, and human oversight to keep the system operating safely and within its intended scope at all times.

Red teaming involves the systematic testing of the system prior to deployment, including the use of chatbots that simulate patients matching the user base’s demographics and psychological traits. Any vulnerabilities identified during this process are resolved before the system goes live, ensuring it has been stress-tested against realistic misuse or unsafe behaviours.

Filtering of dangerous messages provides a real-time automated safeguard at both the input and output stages. User messages are screened by an LLM-based classifier before being processed, intercepting any content that appears to solicit clinical advice or otherwise falls outside the system's intended use. An equivalent filter is applied to all outputs before they are returned to the user. To reduce susceptibility to manipulation, the input filter operates with restricted access to the broader

conversational history, limiting the risk of poisoning attacks.

Classification of user messages for human monitoring extends the automated filtering layer by flagging conversations that exhibit patterns warranting closer attention, even when not immediately intercepted. Rather than relying solely on binary pass-or-block decisions, this classification layer surfaces borderline or ambiguous interactions for review, ensuring that edge cases are not silently passed through the system.

Human monitoring via the *Conversational Care* platform provides the final and most direct layer of oversight. All conversation histories are stored and made accessible to the research team through *Conversational Care* records, allowing staff to review interactions in real time or retrospectively. This ensures that even in cases where automated filters have not flagged a conversation, a human reviewer can identify concerns, intervene where necessary, and inform ongoing improvements to the system's safeguards.

For example, in the social prescribing context, link workers do not need the raw conversational data between young people and the agent to be shared with them (Cho 2025), but they still require visibility into interactions to support safeguarding and inform future care. To address both concerns, the platform introduced a conversation summary tab, providing a reduced representation of interactions while preserving user privacy.

4.2 Fairness

The participatory design of *Conversational Care* aimed to produce a system that supports diverse population groups equitably, varying in ethnicity, gender, age, socio-economic status, geography (urban and rural), primary language, and health literacy - including in low-resource settings (Peters et al. 2026).

Inclusive recruitment shaped the design process throughout. Stakeholders spanned different geographies, socio-economic ranges, caregiver ages, and levels of formal education, with particular attention to older adults, both those living alone and those supported by one or multiple carers.

Rather than assuming a single user profile, the platform supports contextual configuration. During initialisation, care teams set the interaction modality (voice message, SMS, WhatsApp, phone call…), the tone of voice, and the thresholds at which the system escalates to human review, adapting the system to specific populations and to users flagged as vulnerable.

Some features are added for specific contexts. In dementia care, for example, agents pace voice interactions more slowly and rely more on text reminders. Care teams shape the agent's behaviour at set-up and upload local documents that ground its responses.

Prior to deployment, interactions and automations are reviewed by project steering groups and PPI cohorts. After deployment, agent behaviour is verified iteratively, and decision thresholds are adjusted where biases are observed.

Formal performance disaggregation by ethnicity, language variety, and disability status, as recommended in STANDING Together (Alderman et al. 2025), has not yet been undertaken.

4.3 Effectiveness

Robust evidence of medical effectiveness requires a pilot followed by a randomised controlled trial (RCT). We have not completed either, and make no causal claims here. What we can describe is which outcomes matter in our deployments, and where we currently have signal.

Three kinds of outcomes are relevant: patient health (e.g., hospitalisation rates), the wellbeing of relatives (especially in chronic illness and dementia), and the cost and burden on the care team (e.g., monitor hours per family). In chronic illness and dementia, these outcomes need at least 12 months of tracking to be meaningful.

Care-team burden is the outcome on which we currently have the most signal. PPI activities, most extensively within the *UK DRI Minder* project on remote dementia care, identified scheduling, form completion, reminders, filling data gaps, and triage as the tasks perceived as most time-consuming and the priority targets for automation.

A feasibility study is in preparation within *the UK DRI Minder* project, comparing usual care with care that includes these automations (Serban et al. 2026).

Finally, effectiveness needs to be evaluated at two levels: the platform, and - most importantly - each deployed agent. Because agents are locally configured, each deployment will need its own outcome specification rather than inheriting one from the platform.

4.4 Generalisability

Generalisability in healthcare usually refers to whether an effect measured in one site or population transfers to another. We have not studied effectiveness, so we make no claim of that kind. What we can speak to is *design generalisability*: whether the platform is architected so that adaptation to new contexts is feasible. The evidence for this is the set of deployments themselves.

Conversational Care has been instantiated across multiple sites and populations: four Peruvian sites (Lima, Iquitos, Huancayo, Tumbes) within *Impact*

Salud, and UK studies covering dementia care and social prescribing in West London area. These span low- and high-resource health systems, urban and rural settings, and care teams ranging from community health workers to clinicians and link workers.

Three design choices made this possible. First, a **white-label architecture** separates the conversational layer, the knowledge base, and the deployment configuration. Agents can be developed in-house or linked from external services: for some - such as partner-developed sensor-based agents - the platform only needs a server address, and the agent code sits elsewhere. A single instance can also host multiple studies in parallel, with a code (e.g., one provided by a GP) routing each participant to the agents for their study. The same engine therefore supports different agents for different use cases. Second, **modular components** – such as recruitment, consent, user management - are built once and reused, since they are common requirements across medical research. The platform builds on Django CMS, which provides established patterns at both the architectural and interaction-design level (Ravindran 2015). Third, the system is **multilingual and multi-channel**: users interact through WhatsApp, web, or phone depending on context, while dialogue data is stored in a standardised format independent of the interaction platform.

All current deployments require human-in-the-loop configurations. There has been a recurring need for balancing user privacy with administrative or clinical oversight.

When a new deployment starts, prompts and retrieval sources are versioned and re-evaluated before each release.

A defined white-labelling process for new adopters is documented at https://conversational-care.ai.

Integration with existing care systems has been demonstrated through the ongoing use cases, but a plug-in version for wider integration remains to be built.

4.5 Usability and Accessibility

Conversational Care has three audiences with different accessibility needs. (1) Patients, carers, and families, who use the AI agents through familiar consumer platforms (WhatsApp, web, SMS, phone). (2) Care teams - monitors, navigators, clinicians – who use the management platform daily. (3) Researchers and teams who decide to use *Conversational* Care for their studies, who configure the system for new deployments.

For *families*, building on platforms they already know works on low-bandwidth networks and basic devices, which matters in low-resource deployments. Multiple modalities (voice, text, phone) accommodate visual impairment and reading fluency. As reported in Section 3, PPI work with people living with dementia shaped specific choices: slower pacing in voice interactions, accent options, and text reminders alongside voice.

For *care teams*, the management interface was developed through iterative co-design from the start, including message-content decisions. It was evaluated through heuristic evaluation and cognitive walkthroughs, with findings feeding back into navigation, missing features, and safety controls.

For researchers and adopting teams, accessibility means clear documentation, structured onboarding, and ongoing support during set-up - part of the white-labelling process described in Section 4.4.

4.6 Human-Centredness

Conversational Care treats autonomy, competence, and relatedness - the three psychological needs central to Self-Determination Theory (SDT) - as core design requirements rather than evaluation afterthoughts. SDT runs as a cross-cutting frame through the other dimensions (referenced in Safety and Usability and Accessibility), informing the design decisions below (Ryan and Deci 2017).

Patients, carers, monitors, and clinicians are involved throughout, through advisory panels, frequent meetings, and design sprints described in Section 3, shaping scope and feature priorities. In the dementia context, people living with dementia and their carers were treated as members of the design team rather than as participants. Agent messages and content are co-designed with end users and iterated during early piloting.

In the social prescribing case, the system supports the human-centred nature of the care model by helping young people develop self-insight and a better understanding of their care journey. Through reflective nudges related to their participation in prescribed activities, the system shifts the emphasis away from link workers telling young people what to do or explaining the impact of an activity. Instead, it supports young people to make sense of the activities themselves and understand how these may relate to their own mental health. This ultimately aims to help them feel more prepared to engage in co-designed care planning with their link workers afterwards (Cho 2025).

A central human-centred decision was to draw a boundary between what the AI handles and what the human handles. The boundary is set by each care team during the set-up phase and iteratively during deployment, balancing informational and administrative work against relational work. In both *Impact Salud* and *UK DRI Minder*, care teams decided to route emotional, ambiguous, and sensitive matters to the human care team. This is

primarily a safety mechanism (Section 4.1), but its further motivation is to protect the patient-carer-monitor relationship rather than replace it.

In projects where *Conversational Care* is used, participation is opt-in, with explicit opt-out at any time and an alternative non-AI care pathway available for those who decline AI-assisted contact.

Link workers, after trying the low-fidelity model of the system, envisioned that it could encourage young people to talk with them more openly. They felt the system might help young people prepare to share their expectations or concerns before appointments, potentially reducing the time needed to build trust and discuss personal matters.

For example, the agent could prompt young people by asking, “Have you spoken to your link worker about this?”. Care relationships take a long time to develop; as one link worker explained, some had been working with young people for a year because "it takes so long to build a relationship and for them to open up."

4.7 Coordination

Conversational Care sits at this coordination layer: a shared workspace where everyone involved in a person's care can find current information and conversations.

Co-design with monitors and link workers in the UK and with care navigators in Peru pointed to two coordination needs the platform addresses.

- All communications between the care team and a family are kept in one ongoing record, so that a monitor taking over a case partway through can quickly see prior concerns, medication notes shared by the family, and previous interventions (Figure 3, Appendix).
- Shared documents such as care plans can be uploaded, updated, and pushed to everyone involved, removing the duplication that arises when each role maintains its own copy.

Hospital-to-primary-care transition support is out of scope. Care teams in different settings handle these transitions differently, and *Conversational Care* is configured per care team rather than imposing a standard protocol.

4.8 Conviviality

One of the risks of automation in care work is that the redistributed tasks erode the vocational meaning of a role – across nurses, community health workers, monitors, and link workers. Are there uses of CAI that would erase the meaning, joy, and sense of reward that comes from caring for someone? Without a systems view of these social dynamics, and a relationship-centred lens, we may fail to anticipate or even identify social ecosystem harms when they arise. We therefore treat each deployment of CAI as a redistribution of tasks and ask, at each design decision, what that redistribution means for the affected stakeholders.

In the *UK DRI Minder* project, automation was directed at repetitive informational burden - answering recurring questions, sending appointment reminders, scheduling - while in-person visits and emotional support were protected as human-only work. This boundary came directly from prioritisation work with monitoring teams and from PPI sessions in which people living with dementia identified where automation risked eroding human connection (Table 1).

In the social prescribing deployment, the agent bridges the gap between biweekly link worker sessions, supporting continuity for young people when link workers are not present. Conversations around prescribed activities also give link workers visibility into how young people make sense of those activities, adding to the link worker role rather than replacing it.

4.9 Continuity

Three layers affect continuity: **technical** (e.g. does information persist and remain retrievable?), **operational** (e.g. does the system reduce handover time?), and **relational** (does it support a sense of ongoing care?). *Conversational Care* provides system-test evidence for the technical layer, partial qualitative signal for the operational layer from PPI sessions, and none for the relational, which would require multi-year deployment data.

The system is designed around persistence rather than episodes. Every agent interaction is recorded against a single longitudinal record per family of a client, so future care teams can see prior context and reference earlier exchanges. This choice responded to the case-management and longitudinal-monitoring requirements raised by UK monitoring teams in the early pilot (Table 1).

Care plans, care-worker notes, and other contextual information can be uploaded directly. Structured summaries let care workers review previous interactions, while families can query the AI assistant about prior communications and decisions.

Full recordings are available for detailed review, and longitudinal data is presented through dashboard visualisations.

These mechanisms support continuity for the family when the monitor changes, and for the monitor when carers change.

5. DISCUSSION

The two-year long RtD process confirmed that **responsible CAI must be designed and evaluated in context, with its constraints, and with real users consulted on their needs and values**. Each of our deployments required different features to be activated, deactivated, or personalised by the relevant stakeholders. Yet several patterns recurred that we argue should shape how responsible CAI is approached in healthcare more broadly.

First, **care teams need to set the boundary on what to automate**. Across all case studies, teams iteratively chose different levels of automation for informational, administrative, and emotional support tasks. In our developments, the boundary can be dynamic depending on the caseload of the specific support team (Section 4.6).

Second, **an alternative non-AI pathway must remain available**. People differ in needs, digital literacy, and living situations, and a poorly bounded system could widen rather than close the care gap. The platform triages effort towards those who need it most without forcing AI contact on those who decline it (Section 4.6).

Third, **every channel of contact** (e.g., alerts, calls, chatbot exchanges, in-person visits) **is part of the same act of caring for one person**, and must be presented together, in context, so that the human and AI team can act on it (Section 4.7).

Fourth, **stakeholders need to understand what the automation can and cannot do**, and care teams need clear explanatory documentation to share with the people they support.

Together, these patterns are mechanisms for **keeping AI augmentative rather than substitutive, and for protecting the relational meaning of care work** (Section 4.8).

The nine criteria cover a wide set of requirements that individual projects rarely have the budget or time to address from scratch. Our response has been to embody the design decisions made across deployments in *Conversational Care*, a white-label platform built as an application framework. New case studies are likely to surface similar patterns, which **motivates us to open-source the platform for other research and non-commercial projects**.

Design research can elicit needs, prototype, and surface trade-offs, but cannot deliver the longitudinal, validated evaluation that healthcare requires. We have not yet applied the System Usability Scale, NASA Task Load Index, Patient Activation Measure, or a WCAG 2.1 audit (Sections 4.5–4.6); relational impact has so far been observed only through qualitative interviews, and multi-year evidence of relational continuity is still missing (Section 4.9). RtD also carries interpretive risk: our analysis rests on design rationale and documentation rather than formal coding or independent validation, so some findings may reflect the design team's interpretation more than generalisable insight. Agents themselves sit outside the platform core: each is linked to a specific user group or study, and several are developed by partner teams rather than by us. The object of validation must therefore be specified per deployment: the platform architecture is what generalises, while each agent needs its own evaluation. We have not systematically tested individual agents' efficacy, their behaviour against foundation-model drift or evolving care guidelines, nor their failure modes under variable data quality (Sections 4.4–4.6). The platform's dependence on multiple components (LLMs, communication APIs, workflow systems) and on local context also makes replication by other organisations non-trivial; the white-labelling process (Section 4.4) is our first mitigation. FHIR integration with electronic health records remains outstanding, and a defined route for care teams to escalate concerns to clinicians is still under development (Section 4.7). Workforce effects, such as job satisfaction, vocational meaning, and retention (Section 4.8), still need to be evidenced over time. These gaps need to be addressed and measured per deployment.

6. CONCLUSION

This paper described a Research through Design account of building Conversational Care, a white-label platform for human-AI coordination in primary and community care, developed across three projects in the UK, Peru and Singapore. The platform operationalises nine responsible-AI criteria as reusable design and software components, lowering the cost for new projects to address them together rather than one at a time.

The RtD process surfaced three tensions any team building responsible CAI will face: between platform standardisation and contextual fit; between automation as task redistribution and as role replacement; and between design evidence and the clinical evidence healthcare ultimately demands. The platform's architecture - fixed engine, reusable modules, and agents that are locally configured or linked from external services - is one way of holding these tensions productively for the HCI community to build on.

Future work should support per-deployment randomised controlled trials of efficacy, longitudinal study of workforce effects, and fuller documentation of how stakeholder insights became specific UX features, in forms other teams can reuse. The platform is open sourced at https://conversational-care.ai to facilitate both.

7. ACKNOWLEDGMENTS

We thank the people who made this work possible across the UK, Peru, and Singapore: care navigators, community health workers, field workers, monitoring teams, clinicians, link workers, carers, young people in social prescribing, and people living with dementia. Without their work, and the real care they bring to the people they support, this platform would not exist. They motivate us daily.

We also thank the Impact Salud and Reflective Companion (RECO) teams for hosting the fieldwork and pilots, and Kate Sangwon Lee (National University of Singapore) for the Singapore stakeholder research.

This work was funded by NIHR Award ID: NIHR150287 - IMPACT: Innovations using Mhealth for People with dementiA and Co-morbidiTies, the UK Dementia Research Institute [award number UK DRI-7203] through UK DRI Ltd, principally funded by the Medical Research Council, and the ICL/NUS Research Collaboration Fund, and the Leverhulme Trust, through the Leverhulme Centre for the Future of Intelligence.

8. CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Da Re: Conceptualisation, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing.

Fonseca: Conceptualisation, Methodology, Software, Writing – review & editing.

Cho: Investigation (social prescribing), Writing – review & editing.

Serban: Investigation (UK DRI Minder), Writing – review & editing.

Sanchez: Investigation (Impact Salud).

Diez-Canseco: Investigation (Impact Salud).

Cuba-Fuentes: Investigation (Impact Salud).

Vaidyanathan: Investigation (UK DRI Minder).

Calvo: Conceptualisation, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition.

9. REFERENCES

Alderman, Joseph E., Joanne Palmer, Elinor Laws, et al. 2025. ‘Tackling Algorithmic Bias and Promoting Transparency in Health Datasets: The STANDING Together Consensus Recommendations’. *The Lancet Digital Health* 7 (1): e64–88. https://doi.org/10.1016/S2589-7500(24)00224-3.

Bonney, Kathryn, Cory Breaux, Catherine Buffington, et al. 2024. ‘The Impact of AI on the Workforce: Tasks versus Jobs?’ *Economics Letters* 244: 111971.

Car, Lorainne Tudor, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, et al. 2020. ‘Conversational Agents in Health Care: Scoping Review and Conceptual Analysis’. *Journal of Medical Internet Research* 22 (8): e17158.

Cho, Minseo. 2025. ‘Designing Conversational Agents to Support Social Prescribing for Young People’. Paper presented at 38th International BCS Human-Computer Interaction Conference. https://doi.org/10.14236/ewic/BCSHCI2025.28.

Cho, Minseo, Martina C. Di Simplicio, and Rafael A. Calvo. 2026. ‘The Role of Generative Summarisation in Agent-Mediated Reflection for Youth Social Prescribing’. *Proceedings of the 8th ACM Conference on Conversational User Interfaces*, July 21, 1–5. https://doi.org/10.1145/3816046.3816262.

Cho, Minseo, Marco Da Re, Tori Simpson, Matthew Harrison, and Rafael A. Calvo. 2025. ‘Conversational AI in Community Care: Preliminary Insights from a Scoping Review’. Paper presented at 38th International BCS Human-Computer Interaction Conference. https://doi.org/10.14236/ewic/BCSHCI2025.53.

Collins, Katherine M., Ilia Sucholutsky, Umang Bhatt, et al. 2024. ‘Building Machines That Learn and Think with People’. *Nature Human Behaviour* 8 (10): 1851–63.

Cook, Darren, Dorian Peters, Laura Moradbakhti, et al. 2024. ‘A Text-Based Conversational Agent for Asthma Support: Mixed-Methods Feasibility Study’. *DIGITAL HEALTH* 10 (January): 20552076241258276. https://doi.org/10.1177/20552076241258276.

Crabtree, Andrew, Mark Rouncefield, and Peter Tolmie. 2012. *Doing Design Ethnography*. Human–Computer Interaction Series. Springer London. https://doi.org/10.1007/978-1-4471-2726-0.

Da Re, Marco, Fernanda Espinoza, Maria Veronica Belón-Hercilla, et al. 2025. ‘Design and Development of a mHealth Dementia Screening Tool in Resource-Limited Settings: Lessons Learned from Designing for and with Community Health Workers in Peru’. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, April 26, 1–21. https://doi.org/10.1145/3706599.3706662.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. ‘GPTs Are GPTs: Labor Market Impact Potential of LLMs’. *Science* 384 (6702): 1306–8.

Espinoza, Fernanda, Darren Cook, Marco Da Re, María Sofia Cuba Fuentes, Chris R. Butler, and

Rafael A. Calvo. 2025. 'Designing AI-Powered Chatbots for Dementia Care in Peru: Stakeholder Engagement and Field Observations'. *Interacting with Computers*, April 1, iwae057. https://doi.org/10.1093/iwc/iwae057.

Espinoza, Fernanda, Marco Da Re, and Rafael A. Calvo. 2024. 'Designing with Community Health Workers in Latin America'. *Interactions* 31 (5): 55–57. https://doi.org/10.1145/3687148.

Fayad, Mohamed, and Douglas C. Schmidt. 1997. 'Object-Oriented Application Frameworks'. *Communications of the ACM* 40 (10): 32–38.

Illich, Ivan, and Anne Lang. 1973. *Tools for Conviviality*.

Knapp, Jake. 2025. *Click: How to Make What People Want*. 1st ed. With John Zeratsky. Avid Reader Press.

Knapp, Jake, John Zeratsky, and Braden Kowitz. 2016. *Sprint: How to Solve Big Problems and Test New Ideas in Just Five Days*. Simon & Schuster.

Lazo-Porras, Maria, Francisco Jose Tateishi-Serruto, Christopher Butler, et al. 2025. 'Assessment of Health System Readiness and Quality of Dementia Services in Peru: Protocol for a Qualitative Study With Stakeholder Interviews and Documentation Review'. *JMIR Research Protocols* 14 (March): e60296. https://doi.org/10.2196/60296.

Meadi, Mehrdad Rahsepar, Tomas Sillekens, Suzanne Metselaar, et al. 2025. 'Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review'. *JMIR Mental Health* 12 (1): e60432.

Milne-Ives, Madison, Caroline De Cock, Ernest Lim, et al. 2020. 'The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review'. *Journal of Medical Internet Research* 22 (10): e20346.

Molden, Helen. 2024. 'AI, Automation and Psychotherapy–A Proposed Model for Losses and Gains in the Automated Therapeutic Encounter'. *European Journal of Psychotherapy & Counselling* 26 (1–2): 48–66.

Nagendran, Myura, Yang Chen, Christopher A. Lovejoy, et al. 2020. 'Artificial Intelligence versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies'. *Bmj* 368.

Nielsen, Jakob, ed. 1994. *Usability Inspection Methods*. Wiley.

Nielsen, Jakob, and Rolf Molich. 1990. 'Heuristic Evaluation of User Interfaces'. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90*, 249–56. https://doi.org/10.1145/97243.97281.

Peters, Dorian, Fernanda Espinoza, Marco Da Re, Guido Ivetta, Luciana Benotti, and Rafael A. Calvo. 2026. 'Towards Culturally-Appropriate AI for Health in the Majority World: Insights from Participatory Work in Latin America'. *AI and Ethics* 6 (4): 345. https://doi.org/10.1007/s43681-026-01196-y.

Peters, Dorian, Karina Vold, Diana Robinson, and Rafael Alejandro Calvo. 2020. 'Responsible AI—Two Frameworks for Ethical Design Practice'. *IEEE Transactions on Technology and Society* 1 (1): 34–47.

Porras, Maria Lazo, Francisco Tateishi Serruto, Christopher R. Butler, et al. 2025. 'The Journey of the Patients with Dementia in the Peruvian Health System'. *Alzheimer's & Dementia* 21 (S6): e103153. https://doi.org/10.1002/alz70860_103153.

Qin, Jiaqi, Yuanfeixue Nan, Zichao Li, and Jingbo Meng. 2025. 'Effectiveness of Communication Competence in AI Conversational Agents for Health: Systematic Review and Meta-Analysis'. *Journal of Medical Internet Research* 27: e76296.

Ravindran, Arun. 2015. *Django Design Patterns and Best Practices*. Packt Publishing Ltd.

Rittel, Horst W. J., and Melvin M. Webber. 1973. 'Dilemmas in a General Theory of Planning'. *Policy Sciences* 4 (2): 155–69. https://doi.org/10.1007/BF01405730.

Ryan, Richard M., and Edward L. Deci. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Publications.

Sadek, Malak, Rafael A. Calvo, and Céline Mougenot. 2023. 'Co-Designing Conversational Agents: A Comprehensive Review and Recommendations for Best Practices'. *Design Studies* 89 (December): 101230. https://doi.org/10.1016/j.destud.2023.101230.

Serban, Alina-Irina, Marco Da Re, Maria Raposo De Lima, et al. 2026. 'Integration of an Automated Conversational Service Incorporating Speech Analysis in Dementia Care Pathways: Protocol for the Staged Development and Feasibility Evaluation (Preprint)'. Preprint, JMIR Research Protocols, April 28. https://doi.org/10.2196/preprints.99729.

Solano-Kamaiko, Ian René, Ariel C. Avgar, Madeline Sterling, Aditya Vashistha, and Nicola Dell. 2026. 'Sharing the Care: Investigating How Conversational AI Might Facilitate Coordination Among Home Care Workers and Family Caregivers'. *Proceedings of the 2026 CHI Conference on Human Factors in Computing*

Systems, April 13, 1–18. https://doi.org/10.1145/3772318.3791116.

Sundareswaran, Venkataraman, and Arunima Sarkar. 2020. *Chatbots Reset: A Framework for Governing Responsible Use of Conversational AI in Healthcare*. World Economic Forum.

Viitanen, Johanna. 2011. 'Contextual Inquiry Method for User-Centred Clinical IT System Design'. *Studies in Health Technology and Informatics* 169: 965–69.

Wang, Liuping, Zhan Zhang, Dakuo Wang, et al. 2023. 'Human-Centered Design and Evaluation of AI-Empowered Clinical Decision Support Systems: A Systematic Review'. *Frontiers in Computer Science* 5: 1187299.

Wieland, Lillian Carol. 2023. 'Relational Reciprocity from Conversational Artificial Intelligence in Psychotherapy'. *The American Journal of Bioethics* 23 (5): 35–37.

Wiens, Jenna, Suchi Saria, Mark Sendak, et al. 2019. 'Do No Harm: A Roadmap for Responsible Machine Learning for Health Care'. *Nature Medicine* 25 (9): 1337–40.

Zeng, Dian, Lorainne Tudor Car, Kamlesh Khunti, et al. 2026. 'PRIMARY-AI: Outcomes-Based Standards to Safeguard Primary Care in the AI Era'. *Nature Medicine*, 1–4.

Zimmerman, John, Jodi Forlizzi, and Shelley Evenson. 2007. 'Research through Design as a Method for Interaction Design Research in HCI'. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 29, 493–502. https://doi.org/10.1145/1240624.1240704.

APPENDIX

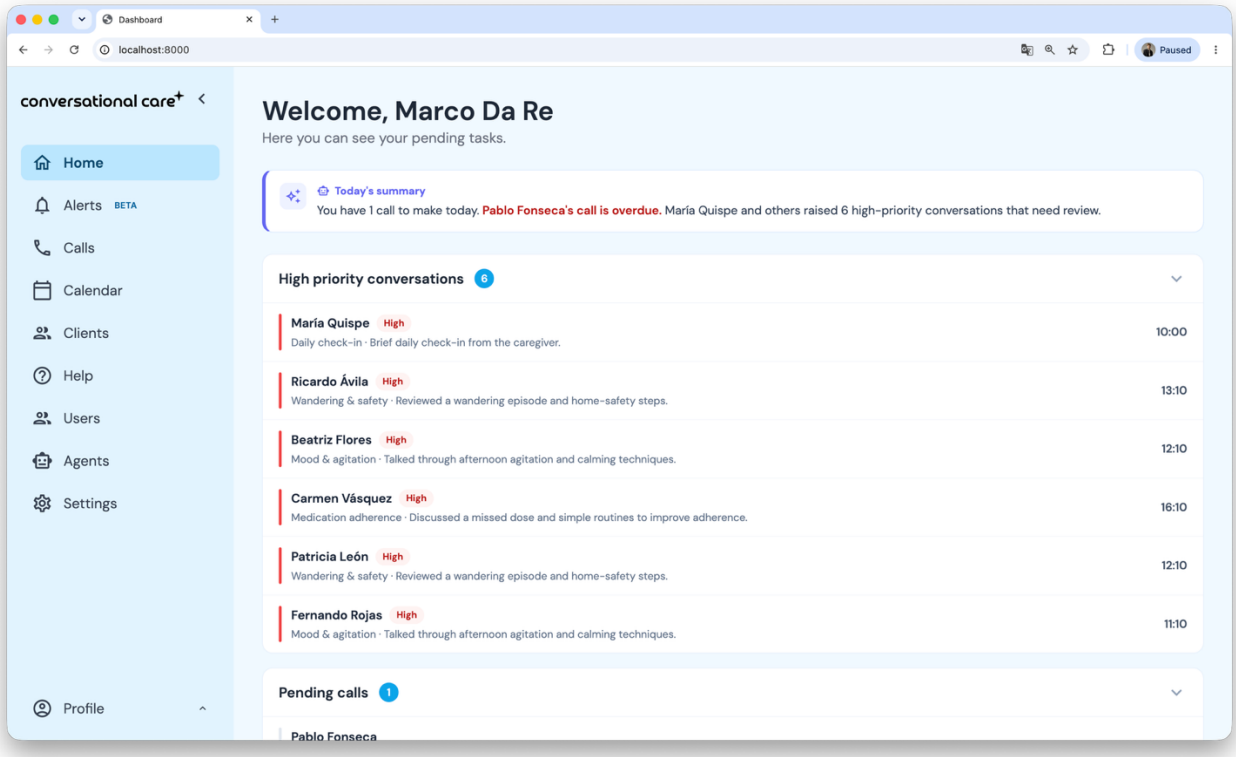


Figure 1: Conversational Care – Dashboard interface

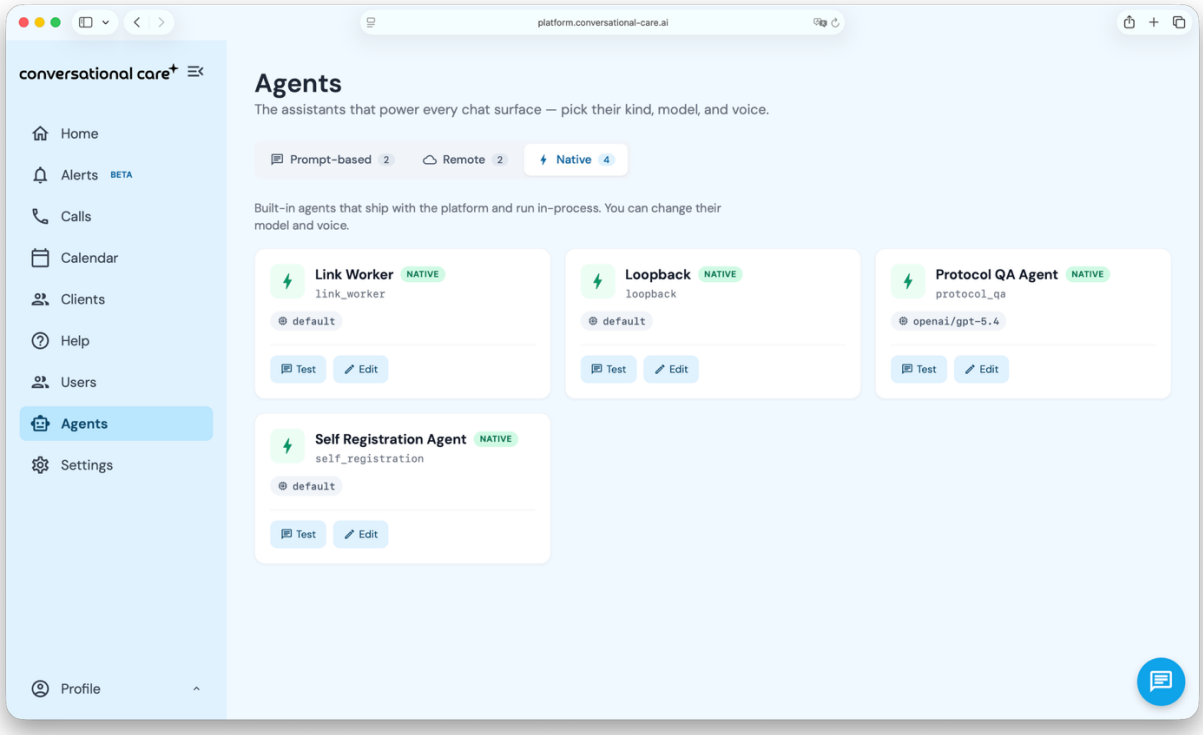


Figure 2: Conversational Care – Pre-defined agents

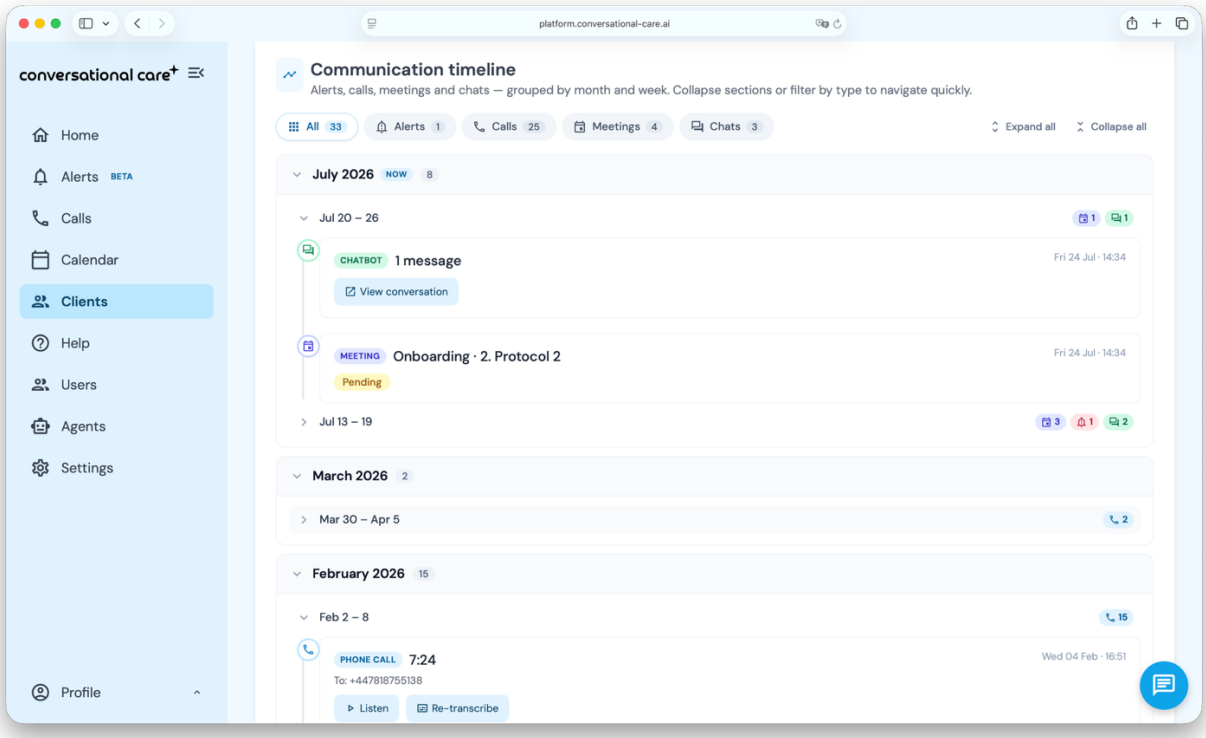


*Figure 3: **Communication Timeline**: A single client record showing alerts, calls, meetings, and chatbot exchanges grouped by month and week.*

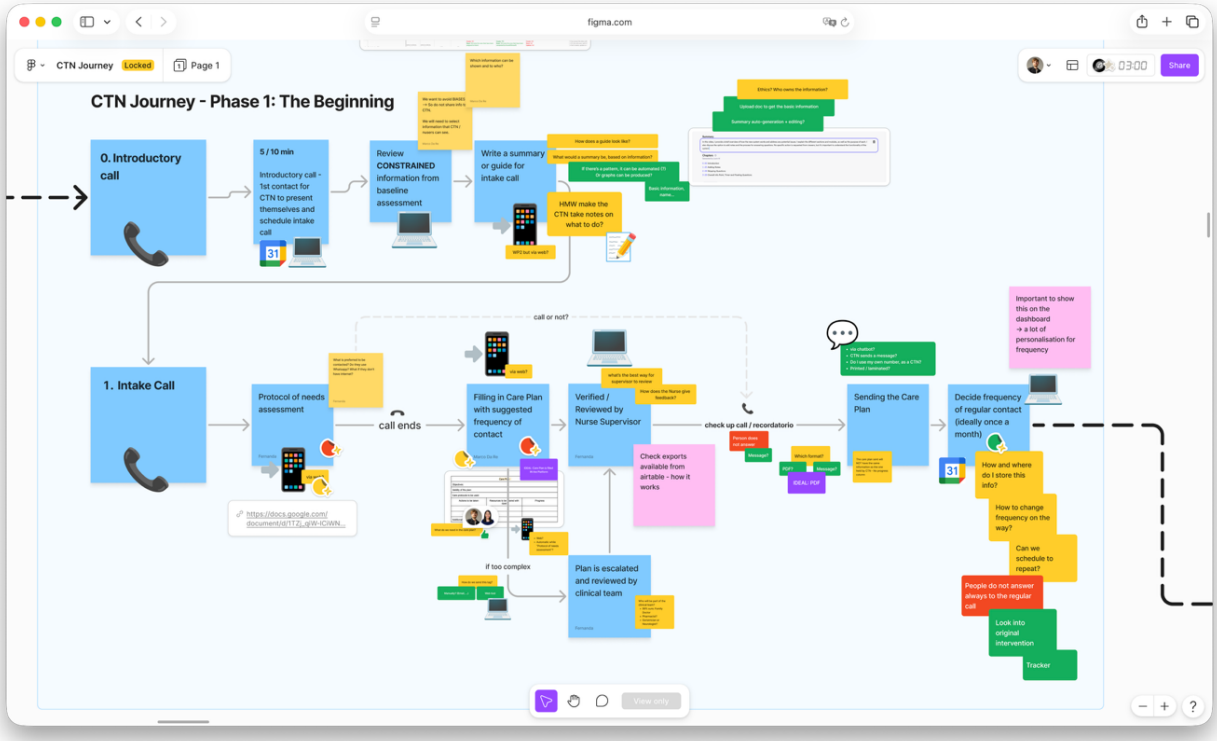


Figure 4: Co-design of a Care Journey Mapping

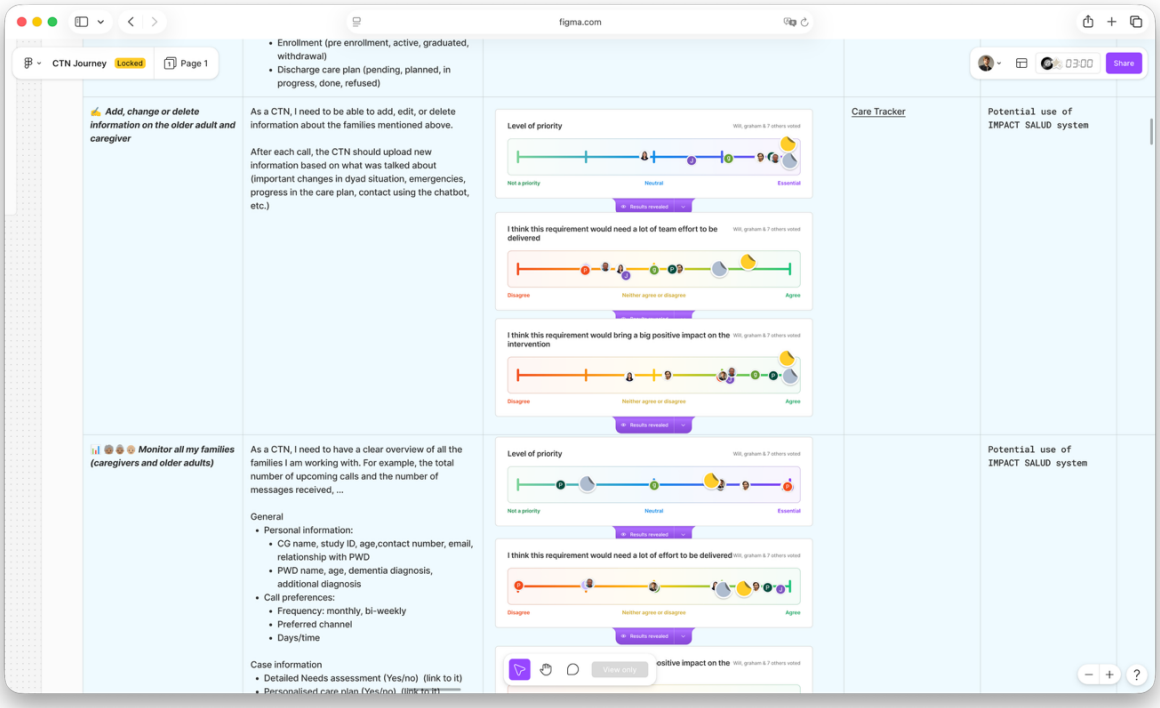


Figure 5: Features Prioritisation

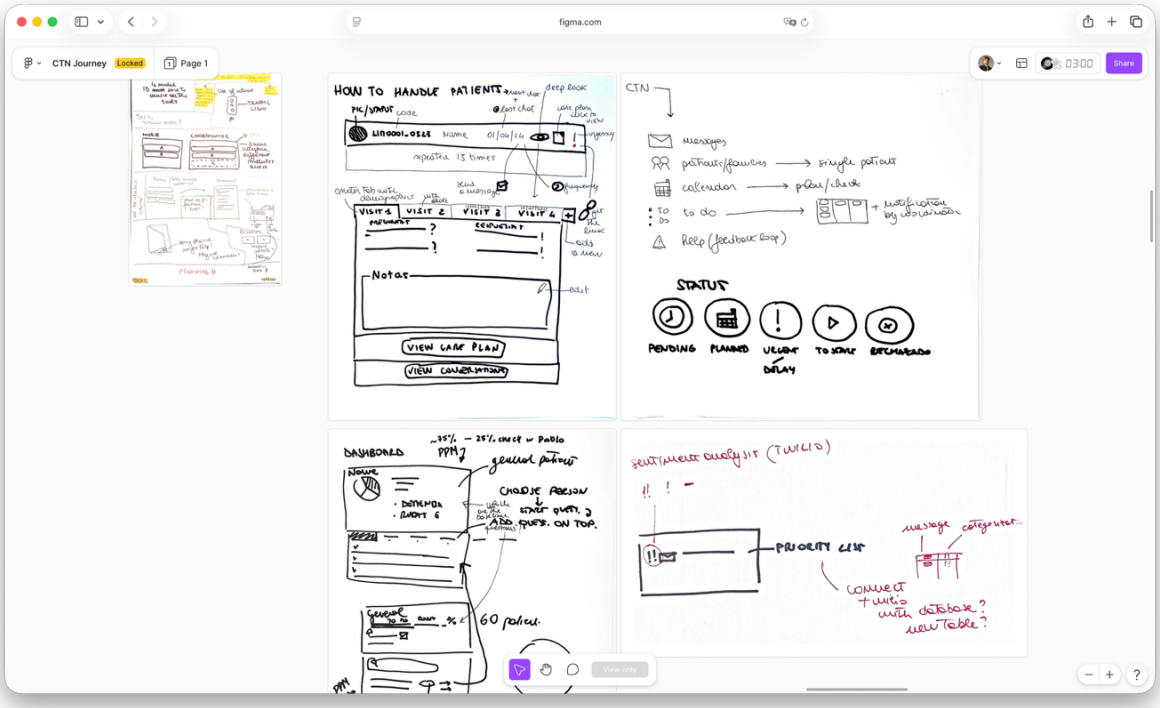


Figure 6: Sketches