

MedIntel: Your Agentic AI Health Companion

For Personal Care and Outbreak Prediction

Ideathon Project Proposal

Team Name: MedIntel

Institution: IIT Kanpur

Members: Namya Jigar Sheth , Neelatmajam Dwivedi

Date: October 22, 2025

Problem Statement

Every week, communities like ours face three persistent healthcare problems—none glamorous, all costly.

🔍 Lack of Precise, Trustworthy, and Local Guidance

The first challenge begins where most health concerns do: with the individual. When people experience minor or recurring symptoms—like a cough, a mild fever, or breathlessness—they need quick, reliable answers. However, the reality is that access to healthcare professionals is often limited. Clinics are crowded, appointments are difficult to secure, and even short consultations can add financial strain.

In these moments, people frequently turn to online searches or social media for advice. But these platforms rarely provide personalized or medically sound guidance. Search engines return generic information without considering context, while social media mixes speculation with misinformation. This leaves individuals overwhelmed, confused, and sometimes misled.

What people truly need is a dependable way to describe their situation naturally—such as “I have a cough for five days, and the air quality is bad today”—and instantly receive evidence-based, localized guidance that integrates environmental factors like AQI, temperature, and current disease trends.

🔄 Break in Continuity of Care

The second, and perhaps more systemic problem, is the loss of continuity in patient care. In today’s healthcare systems, each encounter with a new doctor or specialist often resets the entire patient journey. Medical records are fragmented across hospitals, clinics, and diagnostic centers, and most data sharing still happens manually. As a result, patients must repeatedly recount their medical histories, allergies, test results, and medications.

This repetition not only wastes time but also increases the risk of incomplete or inconsistent information reaching the doctor. For instance, an important past diagnosis, a discontinued drug, or a recurring symptom pattern might be missed simply because it was not transferred properly. The physician then has to rebuild the patient’s profile from scratch—repeating tests that were already done, delaying treatment, and increasing healthcare costs.

This fragmented information flow interrupts what should be a continuous narrative of patient health. It also burdens doctors, who must focus on reconstructing histories rather than providing

precise, forward-looking care. We believe that a unified, intelligent layer that preserves the longitudinal context of each patient could solve this—allowing care to be seamless, informed, and efficient across every provider and setting.

Late Public-Health Signals

The third challenge lies beyond individual care—it’s a community-level problem. Public-health authorities need early warning signals to detect and respond to outbreaks or environmental health risks. The relevant data already exist: local pharmacies record spikes in sales of cough syrups and fever medications, people search online for symptom-related terms, weather systems report humidity and rainfall, and environmental sensors track pollution levels.

Yet, these signals live in isolation. There is no unified infrastructure that fuses them into a coherent early-warning system. As a result, health departments only react after a problem becomes visible—when clinics are already crowded, and resources are stretched thin.

Imagine if these scattered signals could be intelligently integrated. A sudden rise in cough-related pharmacy sales, coupled with worsening air quality and social chatter about sore throats, could alert city health teams days before clinic visits surge. This would enable proactive public-health action—targeted awareness campaigns, local advisories, or preventive resource allocation—well before a full-blown outbreak occurs.

Motivation

Our motivation as a team is both simple and urgent: to close the information and accessibility gaps that fragment modern healthcare. Every person, doctor, and public-health team operates within their own sphere—each with valuable knowledge, but with limited connection between them. We believe this separation is not due to a lack of intent, but a lack of intelligent systems that can integrate these layers seamlessly.

As a team, we asked ourselves: *Why should a citizen’s health guidance, a doctor’s insight, and a city’s early warning system exist in isolation—when all three are part of the same health story?* The problem is not the absence of data; it is the absence of context, continuity, and intelligent fusion.

We envision a world where these boundaries dissolve through technology—where data flows safely, responsibly, and meaningfully between the individual, the clinician, and the community. Our goal is to create a **trustworthy, locally aware, and context-rich ecosystem** that acts as a bridge between personal health experiences, clinical insights, and environmental signals.

Imagine a system that understands when a person describes, “I’ve had a cough for five days, and the air feels bad today,” and instantly correlates it with current air quality data, recent regional disease patterns, and basic preventive recommendations. Now imagine that same data—anonimized and aggregated—helping local health authorities detect a rising trend in respiratory symptoms days before it becomes an outbreak. This is the kind of **connective intelligence** we aim to build.

Our solution revolves around creating an integrated layer that combines three pillars of healthcare intelligence:

- **Individual-level insight** — an AI-driven conversational interface that offers personalized, evidence-based, and context-aware guidance.
- **Clinical continuity** — a secure, longitudinal record that helps doctors instantly reconstruct patient context across visits and providers.
- **Community foresight** — a data fusion engine that aggregates environmental and behavioral signals into real-time public-health alerts.

By connecting these three layers, we aim to transform healthcare from **reactive** to **proactive**—from fragmented to continuous, and from generalized to truly **local and contextual**. Our ambition is not just to build a tool, but to build **trust**—trust that citizens can rely on instant, accurate guidance; that doctors can see the whole patient faster; and that public-health teams can act before risks escalate.

Ultimately, our motivation stems from a shared belief: that better health decisions don't come from more data alone, but from systems that make that data **coherent, compassionate, and actionable**.

Our Solution: *MedIntel*

MedIntel is not just a chatbot — it's an agentic health ecosystem combining:

1. Dual System Design

Our MVP combines:

- **Agentic RAG Health Chatbot:** A safety-first copilot that answers health and environment questions with citations, generates doctor-ready summaries under explicit user consent, and stores them in a secure, portable personal vault.
- **City-Scale Disease Signal:** A lightweight, explainable model that predicts the Top-5 likely diseases for a city or region based on local weather, air quality, crowd symptoms, and pharmacy data.

2. Mutual Reinforcement

The two halves reinforce each other:

- The chatbot (with consent) gathers structured symptom data that strengthens the city's predictive model.
- The city model provides contextual local risk information that makes the chatbot's answers timely and relevant.

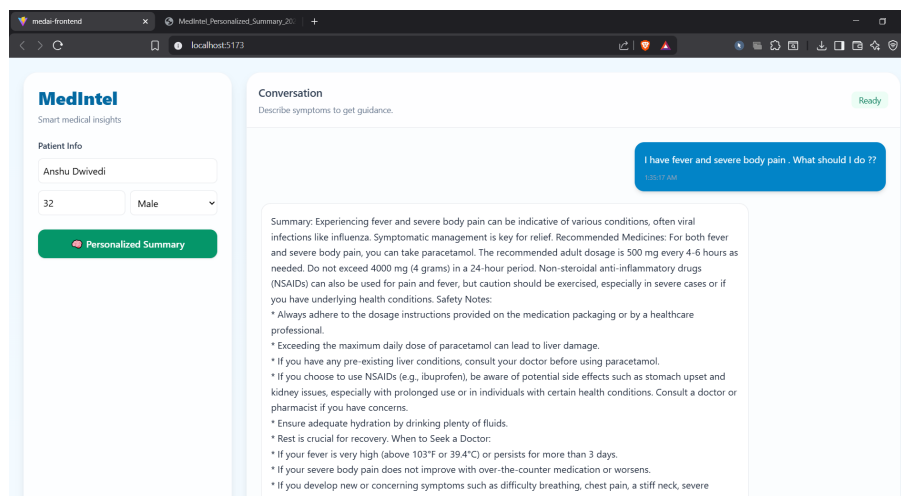


Figure 1: MedIntel System In Action

The 2 Modules of MedIntel

1. AI Medical Chatbot

- Powered by n8n and RAG for context-aware medical reasoning.
- Provides cited, evidence-based answers — no random internet myths.
- Incorporates user medical history (with consent) for safer recommendations.

2. Personal Health Record Manager

- Generates AI-summarized reports for smooth doctor-to-doctor handovers.
- Ensures data security with AES-256 encryption and HIPAA compliance.

Why We Chose Agentic RAG

We deliberately chose the chatbot problem because citizens do not need speculation; they need grounded, local, and cautious guidance. Before committing to Agentic RAG (Retrieve–Act–Generate), we built small proofs of concept for the three dominant chatbot families and pressure-tested them against safety, adaptability, explainability, and maintenance effort.

Alternatives We Rejected (Side-by-Side) Generative-only LLMs (no retrieval).

Strengths: lightning-fast to prototype; fluent answers.

Weaknesses: prone to hallucination, cannot cite sources, hard to keep up-to-date, and risky near clinical topics. With no retrieval constraint, the same question can yield materially different advice across runs. For health + environment, this safety profile is unacceptable.

Intent/FAQ bots (rules, flows, and forms).

Strengths: reliable inside a known script; easy to certify narrow flows; predictable latency.

Weaknesses: brittle outside intent coverage; poor at “unknown unknowns”; cannot integrate dynamic signals (e.g., AQI, heat index) without massive rule growth. As new diseases/advisories appear, maintenance balloons. Citizens ask open-ended questions; rigid flows are the wrong contract.

Plain RAG (retrieve-then-answer once).

Strengths: substantially better grounding; citations possible; updating the KB updates the bot.

Weaknesses: the single-turn pattern cannot plan or call tools, and it struggles with multi-step reasoning: fetch AQI, check a triage list, reconcile two sources, then answer. It can over-quote irrelevant chunks or give premature, under-qualified advice.

Why Agentic RAG is best here. Agentic RAG adds a thin layer of deliberation and tooling above RAG. Our controller plans sub-tasks, calls the retriever, calls external tools (AQI/heat APIs, triage checklists), cross-checks the draft, and only then answers—with strict citations. The agent can also decline gracefully (“insufficient grounded evidence”) or escalate (“red-flag symptoms—seek urgent care”). This yields the safety of grounding, the adaptability of tool use, and the humility to say “I don’t know” when appropriate. In health + environment, where context changes daily and safety margins matter, this is the most defensible architecture.

Why WHO’s data as the Knowledge Base

For ****Version 1.0****, we grounded the chatbot in a ****foundational knowledge base**** comprising *Harrison’s Principles of Internal Medicine*—the same text physicians use—augmented by WHO, CDC, and NHS guidelines and local environmental advisories. This initial selection ensures a strong baseline of clinical reliability, consistent terminology, and trust. We plan to iteratively

expand this knowledge base with additional validated data sources to continuously improve the model's output accuracy and breadth.

To preserve semantics, we use hierarchical chunking ($\sim 1,200$ characters, 200 overlap) with metadata tags: `source`, `edition`, `chapter`, `section`, `url?`, `hash`. Retrieval can be filtered (e.g., endocrine \rightarrow diabetes) and citations are exact. Why this matters: the chatbot is literally reading what clinicians read; alignment reduces conceptual drift and improves trust at the point of answer.

Implementation (Software, In Depth)

High-Level Architecture

- **Runtime:** `n8n` workflow engine (Docker) with `QUEUE_MODE` enabled (Redis) for reliability and back-pressure.
- **Entry points (Triggers):**
 - **Webhook (POST)** `/webhook/midintel/chat` \rightarrow *chatturnsfromapps/IVR*.
 - **Webhook (GET)** `/webhook/midintel/health` \rightarrow *livenessprobe*.
 - **Cron** \rightarrow *scheduledingestionofguidelinedocumentsandoutbreakfeeds*.
- **Core graph (Chat):** `Webhook` \rightarrow Guardrail (IF) \rightarrow Simple Memory \rightarrow AI Agent (Tools: `Simple Vector Store`, `HTTP tools`) \rightarrow Respond to Webhook.
- **Knowledge graph (Ingestion):** `Cron` \rightarrow HTTP Request (docs) \rightarrow Default Data Loader \rightarrow Recursive Character Text Splitter \rightarrow Embeddings (Gemini) \rightarrow Simple Vector Store.
- **Observability:** central `request_id`, execution logs, error sub-workflow, Slack/Email alerts, metrics panel (p95 latency, success rate).
- **Security:** HMAC verification for inbound webhooks, per-route API key, PII redaction node, encryption at rest (`n8n` DB), access tokens in vault.

Knowledge Ingestion and Indexing (`n8n`)

- **Loaders:** HTTP Request nodes fetch guideline PDFs/HTML; Default Data Loader produces document objects with `title`, `url`, `rawText`.
- **Cleaning & structure:** a Function node strips page furniture, normalizes headings, and injects a compact header into each document (`source`, `year`, `section`) before chunking.
- **Hierarchical chunking:** Recursive Character Text Splitter with `chunk.size=1200`, `overlap=200`, word-boundary safe.
- **Metadata:** each chunk carries a JSON `{source, year, section, url, sha256}` field used for filters and deterministic citations.
- **Embeddings & persistence:**
 - **Embeddings:** Embeddings Google Gemini node returns vectors for chunks.
 - **Store:** Simple Vector Store (`n8n`'s built-in) with namespaces by `source_year`; exposed to the AI Agent as a Tool.

- **Verification harness:** post-ingest Function node issues test queries (e.g., “diagnosis criteria ...”) against the store; if $P@k < 0.8$, route to **Alert** branch.
- **Scheduling:** Cron runs nightly; IF node updates only when ETag/Last-Modified differs.

Agent Flow (n8n Chat Graph)

1. Webhook (POST)

Path: /webhook/midintel/chat (Production URL).

Expected Body (JSON):

```
{
  "sessionId": "user_123",
  "message": "I have fever",
  "locale": "en-IN",
  "geo": {"lat": 26.9, "lon": 75.7},
  "client": {"app": "midintel-web", "version": "1.4.0"}
}
```

Guards: HMAC header **x-signature** validated in a Function node.

2. Guardrail (IF)

Reject empty/oversize inputs and prompt-injection patterns.

Rules (regex/examples): tokens > 2,048 →413; contains “ignore previous” →sanitize.

3. Simple Memory

Session ID mode: Connected Chat Trigger Node.

Session Key From Previous Node:

```
{{ $json.sessionId || $json.headers["x-forwarded-for"] || "default_session" }}
```

Context Window Length: 20.

Stored fields: role, content, time, request_id.

4. AI Agent (Chat Mode)

Model: Google Gemini Chat Model node (or drop-in OpenAI compatible).

Tools connected:

- **Simple Vector Store** (tool name: `medical_kb.search`) with filters: `source` (allowlist), `year` \geq latest-1, `section` in ["Diagnosis", "Treatment", "Prevention"].
- **HTTP Request (Env)** (tool name: `env.aqi`) querying AQI/heat index by geohash; memoized via **Cache Map** (key: `geo+hour`).

Inputs (mapped):

- **system_message:** strict, safety-first instructions (citations required; escalate red flags; no diagnostics).
- **input:** `json.message`
- **chat_history:** from Simple Memory.

Output schema (enforced downstream):

```
{
  "summary": "...",
  "guidance": { "steps": ["...", "..."] },
  "citations": [{ "id": 1, "source": "...", "url": "..."}],
  "safety": { "triage": "none|urgent", "reason": "..."}
}
```

5. Post-Processor (Function)

Validates the agent's JSON (shape, citation presence). If invalid, one retry with a sharper instruction; on failure, return cautious fallback with `"safety.triage": "review"`.

6. Respond to Webhook

Returns status 200 with compact JSON and `request_id`. SLA headers include processing time and cache hits.

7. Error Branch (Sub-Workflow)

Any thrown error is routed via **Error Trigger** to a dedicated workflow that logs context, notifies Slack/Email, and writes a failure record (`request_id`, payload hash, stack).

Prompting and Safety

- **Formatting contract:** AI output must be valid JSON (see schema above). A **Function** node hard-parses and rejects free-text.
- **Safety short-circuit:** If the guardrail or agent flags `safety.triage="urgent"`, the response omits general advice and returns a short triage message with emergency signs checklist.

Deployment & Configuration

- **Activation:** workflows toggled Active and invoked via production webhook URLs (not test URLs). Health checks on `/webhook/midintel/health`.
- **Secrets:** model keys and HMAC secrets stored in n8n credentials vault; never hard-coded.
- **Versioning:** export workflows as JSON; commit to VCS with semantic tags (`midintel-chat@1.3.2`, `midintel-ingest@0.9.0`); changelogs enforced in PR.

Concrete Node Settings (Field-by-Field)

Webhook (POST)

- Path: `/webhook/midintel/chat`; Methods: POST; Response Mode: Last Node.
- Binary data: `false`; Raw Body: JSON.
- WebHook Endpoint: `https://iatrochemical-lated-lera.ngrok-free.dev/webhook/e4d3098b-ce6e`

Simple Memory

- Session ID: Connected Chat Trigger Node.
- Session Key From Previous Node:

```
{{ $json.sessionId || $json.headers["x-forwarded-for"] || "default_session" }}
```

- Context Window Length: 20.

AI Agent (Chat Mode)

- Model: Google Gemini Chat Model.
- System Message: safety-first, citation-required, JSON-only.
- Tools: Simple Vector Store (namespace = `medical_kb`), HTTP Request (AQI) with cache.
- Inputs: `input = json.message; chat_history from Simple Memory`.

Simple Vector Store (Tool)

- Search: cosine; Top-k (default): 8; Filters: `year >= now-1, source in allowlist`.
- Return: snippets + metadata (`source, year, section, url, sha256`).

Respond to Webhook

- Status: 200; Body: validated JSON with `request_id, summary, guidance, citations, safety`.

Example Requests & Responses

Request (Chat)

POST /webhook/midintel/chat

```
{
  "sessionId": "user_123",
  "message": "I have a mild fever and sore throat. What should I do?",
  "geo": {"lat": 26.9, "lon": 75.7}
}
```

Response

200 OK

```
{
  "request_id": "b1c0...d9",
  "summary": "...",
  "guidance": { "steps": ["...", "..."] },
  "citations": [{"id": 1, "source": "...", "url": "..."}],
  "safety": {"triage": "none", "reason": ""}
}
```

Why This Split Matters

The separation between ingestion (scheduled, verifiable) and chat (real-time, guarded) yields reproducibility (each node is testable), observability (per-node latency and retrieval fingerprints), and safety (short-circuit on red flags, strict JSON contract).

Why This Matters

This modular, verifiable architecture ensures:

- Safety — all outputs are grounded in medical literature.
- Adaptability — context-aware and locally relevant.
- Trust — transparent citations.

Expected Impact

Patients

Patients benefit significantly from MedIntel through reliable, context-aware, and evidence-based medical guidance. For example, a diabetic patient might receive alerts if their blood sugar patterns indicate a potential risk of hyperglycemia. Furthermore, MedIntel improves health literacy by providing summarized, easy-to-understand advice, such as step-by-step guidance on home hydration, exercise, and diet during heatwaves.

Doctors

For doctors, MedIntel provides AI-curated patient summaries, allowing them to receive concise, structured reports instead of raw data. A clinician, for instance, could review a patient's past six months of lab results and medications on a single page instead of sorting through multiple PDFs. It also serves as a decision support tool, suggesting relevant guidelines and alerting for potential drug interactions, like flagging issues while prescribing antibiotics for elderly patients.

Society

It also fosters improved health literacy by granting wide access to verified, summarized health information, which can be used for neighborhood campaigns explaining proper hygiene practices during flu season. These interventions lead to collective health benefits, resulting in lower disease spread and hospitalization rates.

Impact and Market Projection

Our solution addresses critical gaps in healthcare by enhancing access, improving continuity of care, and enabling proactive public health responses. Below, we present the societal impact and market projections for each component of our solution.

AI Medical Chatbot: Enhancing Access and Efficiency

Societal Impact:

- **Improved Access to Healthcare:** AI chatbots provide 24/7 access to medical information, reducing the burden on healthcare facilities and enabling timely guidance, especially in underserved regions.
- **Cost Reduction:** Automating routine inquiries and preliminary assessments can save the healthcare industry globally around \$3.6 billion by 2025.
- **Enhanced Patient Engagement:** Chatbots deliver personalized health advice, reminders, and educational content, promoting proactive patient participation in care.

Market Projection:

- **Global Market Growth:** The AI in healthcare market is projected to grow from USD 26.6 billion in 2024 to USD 187.7 billion by 2030 .
- **Adoption Rates:** Widespread adoption could save up to 10% of U.S. healthcare costs annually (\$360 billion) .

📌 Personal Health Summarizer: (PHS) Generator and Summarizer: Facilitating Continuity of Care

Societal Impact:

- **Improved Health Outcomes:** Comprehensive health summaries enable providers to make informed decisions and improve patient outcomes .
- **Enhanced Patient Empowerment:** PHSs give patients control over their health information, increasing engagement and health literacy.
- **Reduction in Medical Errors:** Consolidated information reduces errors from incomplete or inaccurate histories.

Market Projection:

- **Adoption Rates:** Increasing recognition of PHS benefits is driving adoption worldwide.
- **Economic Impact:** PHSs reduce duplicate testing, hospital readmissions, and improve healthcare delivery efficiency.

Future Plans for the upcoming versions — Neighborhood-Aware Prediction & Personalized Doctor Interaction

Vision

We will extend MedIntel to (A) leverage outbreak signals from nearby cities and mobility networks to improve early warning at a target location; and (B) add secure, one-on-one personalized doctor-patient interactions inside the chatbot — combining public health foresight with clinical continuity. This makes MedIntel more accurate, more actionable, and far more useful for real people and health systems.

A. Neighborhood-aware Outbreak Prediction (Why and How)

Motivation. Infectious events rarely respect administrative borders: an outbreak in City A often spreads to nearby City B via commuting, markets, and festivals. Observing signals from neighboring areas provides early indicators that local time-series alone might miss.

Core ideas / components:

- **Spatial pooling:** incorporate data from geographic neighbours (e.g., within 50–200 km) with distance / connectivity weighting.
- **Mobility-informed weights:** use mobility proxies (commuting matrices, aggregated mobility reports) to weight influence of each neighbour rather than pure geographic distance.
- **{Spatio-temporal model:** upgrade the LSTM to a spatio-temporal architecture (options below) so the model learns how events propagate across space and time.
- **Transfer learning & fine-tuning:** pretrain on regions with abundant outbreaks, then fine-tune on data-sparse target cities using neighbor data for improved sample efficiency.
- **Anomaly fusion:** combine local anomaly detectors (sudden spikes in user reports or drug sales) with neighbor trends to reduce false positives.

Modeling options (in increasing sophistication):

1. **Weighted neighbor aggregation (simple, robust):** build neighbor feature vectors by computing a weighted average of neighbor time-series and append to target sequence. Good for MVPs.
2. **Graph-augmented temporal model (powerful):** construct a region graph $G = (V, E)$ where edges are mobility-based; apply a Graph Neural Network (GNN) to propagate signals across G and feed outputs into temporal LSTM layers (GNN+LSTM / ST-GNN).
3. **Spatio-temporal attention:** allow the model to attend to both specific past timesteps and particular neighbor nodes (e.g., attention highlights “last week, City X’s drug sales rose sharply”).
4. **Probabilistic ensembles:** ensemble the LSTM/GNN outputs with rule-based heuristics (e.g., exceedance of sales z-score) and calibrate using validation data.

Key Note

The outbreak model becomes smarter with every user (After full consent). Your data (completely anonymized) trains a system that can help your community stay safer.

Real-World Advantage

Unlike static government datasets, MedIntel’s model evolves daily using fresh user data — an adaptive early-warning system.

Practical considerations and data sources:

- **Mobility data:** use Google Community Mobility, Apple Mobility trends, or anonymized telecom aggregates to build $w_{r \leftarrow j}$.
- **Granularity alignment:** harmonize neighbor signals to the same weekly cadence and region ID scheme.
- **Cold-start cities:** for new target cities with sparse user data, rely more on neighbor signals and gradually shift weight toward local signals as local data accrues.
- **Evaluation:** compare pure-local vs neighbor-augmented models on historical outbreaks using lead-time gain (how many extra days earlier we detect an outbreak) and PR-AUC.

Benefits:

- Earlier detection via upstream signals.
- Fewer missed outbreaks in border or commuter towns.
- More robust predictions for data-sparse regions.

B. One-on-One Personalized Doctor–Patient Interaction inside the Chatbot

Motivation. A connected, contextual conversation with a clinician (via the chatbot) closes the loop: users get personalized advice and doctors get structured context — improving care continuity and trust.

Feature set (what the interaction provides):

- **Secure messaging and notes:** encrypted chat between user and clinician with attachments (reports, images).
- **Smart summaries:** AI summarizes a user's PHR into a concise one-page clinical brief for the doctor (diagnoses, meds, recent labs, allergies).
- **Symptom triage + decision support:** chatbot performs initial triage (red/yellow/green) and supplies structured prompts to clinicians (e.g., "check CRP; consider stool culture").
- **Appointment follow-up workflow:** schedule booking, reminders, and automated follow-up checklists (symptom trackers).
- **Consented record sharing:** one-click share of selected records with a clinician for a limited time window.
- **Contextual model assistance:** when clinician is writing a note, the AI suggests coded diagnoses, drug interaction checks, and relevant guidelines (e.g., CDC/WHO).

Impact Summary

By combining neighborhood signals and personalized clinician interactions, MedIntel will detect outbreaks earlier and convert early warnings into timely clinical actions — improving both public health response and individual care.

Closing note

This future direction turns MedIntel from a forecasting tool into a living health platform: it senses risk in the neighborhood and connects users seamlessly with clinicians — so warnings become timely actions and data becomes trusted, shared health intelligence. Provide a short tagline for the project

Tagline

"Because your health deserves more than a Google search — it deserves an AI sidekick that actually studied medicine."

End of Proposal — Stay Healthy, Stay Curious!