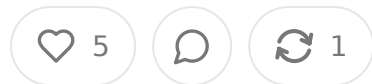


The PCP as Specialist: How AI and Virtual Consults Will Collapse the Referral Economy and Create a New Category of Primary Care Company

APR 12, 2026



Share

Abstract

This essay outlines a practical business plan for a practicing primary care physician (PCP) who wants to build a next-generation AI clinical care company. The core thesis is that AI will allow PCPs to manage conditions traditionally referred out to specialists, specifically in cardiology, dermatology, endocrinology, nephrology, and other chronic disease verticals, by keeping the specialist in the loop virtually before a referral is ever made.

- The problem: roughly 9% of all ambulatory PCP visits result in a specialist referral, 20-30% of which are for routine clinical issues that could be safely managed in primary care. The average downstream cost of a single specialist referral is roughly \$965. Over 100 million specialist referrals are issued annually in the US, and only about half are ever completed.
- The model: an AI-powered clinical decision support layer embedded in the PCP workflow that triages, recommends, and connects the PCP to a specialist via asynchronous eConsult only when needed, allowing the PCP to manage more conditions in-house.
- The stakeholder value: PCPs earn more per patient and practice at the top of the market. Specialists focus on truly complex cases. AI vendors get distribution. Health systems get a new category of primary care company.

plans reduce total cost of care. Investors get a capital-efficient SaaS-plus-service model. Patients stop falling through referral cracks.

- The founder path: how a practicing PCP with no tech or business background bootstrap this from a single-site pilot to a scalable company, with specific guidance on team building, fundraising, regulatory navigation, and go-to-market.

Table of Contents

1. The referral problem nobody talks about enough
2. The thesis: AI as the new specialist triage layer
3. What the product actually looks like
4. The specialist-in-the-loop model
5. Unit economics and the business model
6. What each stakeholder gets out of this
7. The PCP founder playbook
8. Regulatory and liability considerations
9. Go-to-market and early traction
10. Why investors should care
11. Where this breaks and what to watch

The referral problem nobody talks about enough

Here is the dirty secret of American primary care. The typical PCP interacts with 200 different specialists in a given year. That number is wild if you sit with it for a second. The referral rate from PCP visits roughly doubled between 1999 and 2009.

going from about 4.8% to 9.3% of all visits, and by most accounts it has only continued climbing since. One out of every ten visits to a primary care doc now results in a referral to someone else. And when those referrals go out, the loop almost never cleans up. Research on large health systems has shown that a huge proportion of referrals never result in a completed specialist appointment with results flowing to the PCP. Out of those 100-plus million specialist referrals issued annually in the US, roughly half are never completed. Patients get lost. Notes don't get sent back. Nobody follows up.

This is a structural problem, not a competence problem. PCPs are stuck in 15-minute visit windows. They know that if a patient presents with something outside their comfort zone, the safest move (legally and clinically) is to punt to a specialist. The specialist gets the referral, maybe sees the patient 6-8 weeks later, runs a bunch of tests, and frequently concludes that the PCP could have managed this themselves with a minor medication adjustment or a watchful waiting plan. Meanwhile the patient has burned a day off work, paid a copay, sat in another waiting room, and got anxious about what might be wrong. Then nobody tells the PCP what happened, and the cycle repeats.

The cost side is real. The average total downstream cost from a single specialist referral runs about \$965 when you factor in office visits, labs, imaging, and procedures. Multiply that across the system and the numbers get staggering fast. About one in three hospital discharges also generates a specialist referral. Two out of three Medicare beneficiaries have two or more chronic conditions requiring care from multiple specialties. And the system is actually getting worse, not better.

UnitedHealthcare rolled out sweeping new referral requirements for its Medical Advantage HMO plans starting January 1, 2026, requiring PCPs to submit referrals to UHC before specialist visits. That is UHC essentially saying: we think too many of these referrals are unnecessary, and we want the PCP to be the gatekeeper again. The payer world is screaming for someone to fix this.

The most interesting data point in all of this comes from the eConsult literature. Studies have shown that eConsults (asynchronous provider-to-provider consultations where a PCP sends a clinical question to a specialist electronically) can reduce the

need for face-to-face specialist visits by up to 70%. At Geisinger, adoption of the Ask-a-Doc eConsult system reduced specialist office visits by 74% in the first month. A randomized study of Medicaid patients found that total costs declined by \$651 per patient in the eConsult group compared to traditional referrals. PCPs rated satisfaction with these services above 4 out of 5, and 78% of patients who experienced an eConsult said they would prefer it over a face-to-face referral in the future.

So the demand signal is clear. Payers want fewer unnecessary referrals. Patients want faster answers. PCPs want to feel more competent and less like a routing layer. Specialists are drowning in cases that don't need their full attention. The question is what happens when you layer modern AI on top of this already proven eConsult model.

The thesis: AI as the new specialist triage layer

The eConsult model works, but it has a bottleneck. The specialist still has to read the case, review the attached documents, and type up a recommendation. That takes time. Even at an average turnaround of two days (which is the Ontario eConsult average across nearly 100,000 cases), that is still friction. And the quality of the consult depends entirely on how well the PCP framed the question and how much context they attached. Garbage in, garbage out.

AI changes the equation in three ways. First, it can help the PCP frame the clinical question properly before it ever reaches a specialist. Think of it as a pre-consult copilot. The PCP describes the presentation, uploads relevant labs, images, or notes, and the AI synthesizes everything into a structured clinical summary with a preliminary differential diagnosis and recommended workup. This alone would reduce the specialist 80% of the cognitive work they currently do on routine eConsults. Second, the AI can triage cases into three buckets: cases the PCP can manage themselves with guideline-based recommendations, cases that need an asynchronous specialist consult, and cases that genuinely need a face-to-face specialist referral. The triage function is where the real value sits because it routes only the right cases

right level of care. Third, the AI can learn from every consult interaction over time, building a knowledge base that gets smarter about which presentations in which contexts with which patient histories actually need specialist input versus which are safely managed in primary care.

This is not about replacing specialists. Nobody credible is arguing that a PCP armed with AI should be doing cardiac catheterizations or managing complex type 1 diabetes in a pregnant patient. The argument is about the 20-30% of referrals that are for routine clinical issues. The patient with mildly elevated TSH who just needs a medication titration plan. The patient with a suspicious skin lesion that a dermatologist could assess from a photograph. The patient with stage 2 CKD whose medication list needs adjustment. The patient with stable atrial fibrillation on appropriate anticoagulation who just needs monitoring. These are cases where a PCP supported by AI clinical decision support and with a specialist available asynchronously for a quick confirmation, can deliver care that is as good as or better than the current referral pathway. Better because it is faster, cheaper, and keeps the patient in a longitudinal relationship with the provider who actually knows them.

What the product actually looks like

At the most basic level, this is a software platform that sits inside or alongside the PCP's existing EHR workflow. The PCP encounters a clinical scenario that would traditionally trigger a referral. Instead of generating a referral order, they open the consult module. They can describe the clinical scenario in natural language (or the system can pull structured data directly from the chart if integrated with the EHR). The AI engine processes the inputs against clinical guidelines, recent literature, and the platform's own accumulated consult history. It generates a structured output that includes a clinical summary, a preliminary assessment, recommended next steps, and a confidence score.

If the AI confidence is high and the recommendation falls within established guidelines for PCP management, the platform presents the treatment plan directly to the PCP with relevant supporting evidence. The PCP reviews, accepts or modifies

plan, and documents the encounter. No specialist involved. If the AI confidence is moderate or the case has complicating factors, it packages the case for asynchronous specialist review. The specialist gets a clean, structured summary with the AI's preliminary assessment and the PCP's specific question, reviews it, and responds with a recommendation. Turnaround target is under 24 hours. If the AI identifies red flags or the case is clearly complex, it recommends a traditional face-to-face referral to the specialist, but even here it pre-populates the referral with structured clinical information so the specialist visit is more productive from minute one.

The technical stack is not as complicated as it might sound. The core is a clinical reasoning engine built on a large language model fine-tuned on medical literature, clinical guidelines, and (critically) real eConsult case data. You need an integration layer to pull data from EHR systems, which realistically means FHIR APIs and partnerships with the major EHR vendors. You need a secure messaging layer for asynchronous specialist consults, which is essentially what companies like RubiconMD and AristaMD have already built. You need a clinical image handling pipeline for dermatology, wound care, and other visually-driven specialties. And you need an analytics layer that tracks outcomes, measures accuracy, and feeds learn loops back into the model.

The key technical differentiator is the triage function. Getting this right is everything. If you over-triage (sending too many cases to specialists), you are just an expensive eConsult platform. If you under-triage (keeping cases in primary care that should have been referred), you have a patient safety problem. The calibration of this triage engine is what the entire company is really about.

The specialist-in-the-loop model

This is the part that makes the whole thing work and also the part that separates it from the dozens of AI clinical decision support tools already on the market. The specialist is not replaced. The specialist is repositioned. Instead of seeing 30 patients a day, 10 of whom could have been managed by the PCP, the specialist reviews 30

asynchronous consults in an hour (most of which are pre-digested by the AI) and 20 patients face-to-face, all of whom actually need their expertise.

The specialist gets paid for eConsult reviews. In the VA system, specialists receive workload credit for eConsults at one of three levels based on time spent. At Mayo Clinic, eConsults are scheduled as 15-minute appointments with visit credit. Several state Medicaid programs now provide a transactional payment for eConsults to either the PCP, the specialist, or both. CMS has been slowly moving toward recognizing interprofessional consultation codes (CPT 99451, 99452) that compensate both providers for eConsult activity. The payment infrastructure is not fully mature, but is getting there, and value-based arrangements accelerate this because the savings from avoided referrals flow directly to whoever is managing total cost of care.

The specialist-in-the-loop model also addresses the biggest objection to AI-assisted primary care, which is liability. When a specialist reviews and co-signs an AI-generated recommendation, you have a documented chain of clinical reasoning that includes specialist input. That is arguably more defensible than the current model where the PCP either refers (and the patient never goes) or manages the condition alone without any specialist input at all. The eConsult creates a record. It shows the PCP asked the right question, the AI synthesized relevant data, the specialist reviewed and agreed (or modified the plan), and the PCP implemented it. From a malpractice perspective, that documentation trail is stronger than what exists in clinical encounters today.

For the specialist personally, this model is actually attractive. The data from eConsult satisfaction studies consistently shows specialists rate these interactions favorably. They get to practice at the top of their license. They stop seeing patients who do not need them. They get paid for cognitive work without the overhead of a face-to-face visit. And their in-person panels become more interesting and more clinically challenging, which is what most specialists actually went into their field to do.

Unit economics and the business model

There are a few ways to monetize this. The cleanest is a per-consult fee charged to the entity managing total cost of care. In a fee-for-service world, that entity is the payer. In a value-based world, it could be the ACO, the health plan, or the provider organization itself. AristaMD and RubiconMD have proven this model works. RubiconMD reports that primary care clinics treating Medicare and Medicaid populations save hundreds of dollars per eConsult on average before even factoring in downstream savings from avoided complications.

The AI layer adds margin. The cost of an AI-assisted triage is pennies per query from a compute standpoint. The cost of a specialist eConsult review is \$25-\$100 depending on complexity. If the AI can resolve 30-40% of cases without any specialist involvement (because they fall squarely within guideline-based PCP management), that is pure margin. The platform collects the per-consult fee but only pays the specialist on the subset that actually needs their input.

A reasonable model for a single PCP practice might look something like this. As a PCP sees 20 patients per day, 250 days per year, so 5,000 patient encounters annually. At a 9% referral rate, that is 450 referrals per year. If the platform can resolve 30% of those without a specialist (135 cases managed by the PCP with AI guidance), route 50% through asynchronous specialist eConsult (225 cases), and only 20% as traditional referrals (90 cases), the value creation is substantial. At \$1,000 average downstream referral costs, avoiding 135 traditional referrals saves \$135,000. Even if the platform charges \$50 per AI-assisted consult and \$75 per specialist eConsult, the total platform cost to the practice or payer is \$23,625 per year. The value is obvious.

For a platform company, the economics improve with scale. The specialist panel is a shared resource across many PCP practices. The AI model improves with volume. Customer acquisition cost in healthcare is brutal (18-24 month sales cycles for hardware systems), so the go-to-market needs to target independent practices and small groups first, then move upmarket with data and outcomes.

Revenue model options include SaaS subscription (flat monthly fee per PCP), per-consult transaction fees, or value-based arrangements where the platform takes

percentage of documented savings. The smartest play is probably a hybrid: a low monthly subscription that covers the AI triage layer, plus a per-consult fee for specialist eConsults, plus a performance bonus tied to referral reduction and outcomes metrics in value-based contracts. That gives you recurring revenue, upside, and alignment with the direction payers and CMS are moving.

What each stakeholder gets out of this

Patients get faster answers. Instead of waiting 6-8 weeks for a specialist appointment (if they even make the appointment at all, which half don't), they get specialist-informed guidance at their PCP visit or within 24 hours. They stay in a relationship with a provider who knows their full history. They avoid the burden of additional appointments, copays, and time off work. Data from eConsult studies shows 78% of patients who experience an eConsult prefer it to a traditional referral.

PCPs get to practice at the top of their license. They stop feeling like a routing slip that just triages patients to other doctors. They build clinical competence over time because every AI-assisted consult is a learning opportunity. They increase their revenue per patient by managing more conditions in-house (billing for the management visit instead of generating a referral that pays them nothing). In value-based contracts, they capture savings from reduced total cost of care. And they reduce their malpractice exposure by creating documented, specialist-informed clinical decisions.

Specialists get their time back. They stop drowning in routine cases. Their face-to-face panels become more clinically interesting. They earn income from eConsult reviews, which can be done from home, on their own schedule, without clinic overhead. They extend their reach to patients they would never have seen otherwise, particularly in rural and underserved areas. And they build reputational equity as an expert network behind a growing platform.

Health plans get what they have been asking for: lower total cost of care without restricting access. The eConsult model reduces unnecessary specialist visits (the biggest cost driver in ambulatory care), catches conditions earlier before they es

to ED visits or hospitalizations, and creates data trails that support quality measurement and risk adjustment.

AI vendors get distribution. The clinical reasoning engine at the center of this platform could be built on top of existing foundation models (Anthropic, OpenAI, Google) with medical fine-tuning, or it could leverage emerging medical-specific models. Either way, the platform provides a channel for AI technology to reach clinical workflows in a way that is validated, supervised, and scalable.

Investors get a company with strong unit economics, a moat that deepens with data, multiple revenue streams, and tailwinds from every major trend in healthcare: value-based care, provider shortage, AI adoption, payer cost pressure, and regulatory support for interprofessional consultation.

The PCP founder playbook

This is the section that matters most for the practicing doc who reads this and thinks “okay, but how do I actually do this when I don’t know how to write code or build a pitch deck?”

Step one is to start with the clinical workflow, not the technology. Before writing a single line of code, spend three months documenting every referral generated in practice. Log the specialty, the clinical question, the outcome (did the patient go to the specialist, did the specialist recommend, could the PCP have managed it). This referral audit is the foundation of everything. It tells you which specialties to target first, what percentage of referrals are potentially avoidable, and what the clinical questions actually look like in practice. This data becomes your pitch to investors and your product requirements doc for engineers.

Step two is to find a technical cofounder or a fractional CTO. This does not mean hiring a full engineering team on day one. It means finding one person, ideally someone with health tech experience, who can build a minimum viable product. MVP is embarrassingly simple: a web form where the PCP enters a clinical question, an AI engine that generates a structured summary and preliminary recommendation,

and a secure messaging channel to a specialist who can review and respond. That's it. No EHR integration. No fancy dashboards. Just the core workflow. You can build this in 8-12 weeks with a small team.

Step three is to recruit 3-5 specialists who are willing to participate as the initial consult panel. These should be docs the PCP already has relationships with, people who are intellectually curious about the model and willing to do eConsult review at a modest per-consult rate during the pilot. Cardiologists and endocrinologists are the first targets because those specialties have the highest volume of routine referrals that could be managed in primary care with guidance.

Step four is to run a pilot in the PCP's own practice for 6-12 months. Track everything: referral rates before and after, time to clinical resolution, patient satisfaction, specialist turnaround time, AI accuracy (how often does the specialist agree with the AI recommendation, and how often do they modify it), and cost data. This pilot generates the clinical evidence and the business case for fundraising.

Step five is to raise a pre-seed round. The PCP founder brings the clinical expertise, the pilot data, and the domain credibility. The technical cofounder brings the product. Together they need \$500K-\$1.5M to hire a small engineering team, expand the specialist panel, and onboard 10-20 additional PCP practices. Target health tech angels and seed funds that understand the eConsult model and the value-based care thesis. The Y Combinator healthcare cohorts in 2025 and 2026 include companies like Locata (AI referral management for primary care) and Clara (AI primary care platform) so there is clearly investor appetite for this category.

Step six is to pursue early payer partnerships. Approach a regional Medicaid managed care plan or a Medicare Advantage plan with the pilot data and propose a value-based arrangement where the platform gets paid a percentage of the savings generated from reduced specialist referrals. This is the fastest path to revenue at scale and it aligns incentives perfectly.

The PCP founder does not need to become a tech CEO. The PCP founder needs the chief clinical officer and the face of the company to the provider community. The day-to-day CEO role can be filled by a healthcare operations executive recruited after the seed round closes. The founder's job is to make sure the product works clinically, the specialist network trusts the platform, and the PCP customers feel like this was built by someone who understands their world. That last part is the unfair advantage that no Silicon Valley team can replicate.

Regulatory and liability considerations

The regulatory picture is less scary than most PCP founders assume, but it requires attention. The platform is not practicing medicine. It is providing clinical decision support. The AI generates recommendations. The PCP makes the clinical decision. The specialist provides consultation. Those are all well-established roles with clear legal frameworks.

FDA regulation of clinical decision support software has been clarified through the 21st Century Cures Act. Software that is intended to support clinical decision-making (rather than replace it) and that enables the clinician to independently review the basis for the recommendation generally falls outside FDA device regulation. The key is that the software must display its reasoning and the underlying data, and the clinician must be able to evaluate the recommendation independently. This means the platform needs to show its work, not just spit out a recommendation as a black box.

HIPAA compliance is table stakes. The platform handles PHI and needs to be compliant with all the usual requirements: encryption in transit and at rest, BAAs with all vendors, access controls, audit logging. Nothing exotic here, but it needs to be done right from day one.

State medical licensing is the tricky one for the specialist eConsult component. A specialist reviewing the case needs to be licensed in the state where the patient is located, or the interaction needs to qualify under an exception. Several states have adopted interstate medical licensure compacts that make this easier. Alternative structuring the specialist input as an interprofessional consultation (provider-to-

provider, not provider-to-patient) can sidestep some of the licensing requirements since the specialist is advising the PCP rather than treating the patient directly.

Malpractice insurance carriers are increasingly familiar with eConsult and telemedicine models. The PCP's existing malpractice policy generally covers clinical decisions made with the aid of decision support tools, and the specialist's policy covers their consultation recommendations. The platform company should carry its own professional liability coverage as well.

Go-to-market and early traction

The go-to-market for a company like this has to be bottom-up, at least initially. Health system sales cycles run 18-24 months and require multiple stakeholder sign-offs. Independent PCP practices can make a buying decision in a week. The initial target should be independent and small-group PCP practices (2-20 providers) in areas with favorable telehealth and eConsult reimbursement policies. There are roughly 200,000 primary care physicians in the US, and a significant percentage of them are still in independent or small-group practice.

The pitch to the PCP is simple: "You know those referrals you send out every day for stuff you could probably manage yourself if you just had someone to bounce it off of? What if you had an AI copilot that helps you manage those cases in-house, with a specialist backing you up asynchronously, and you get to bill for the management instead of generating a referral that earns you nothing?"

Early traction comes from the PCP founder's own network. Every PCP knows other PCPs. The first 20 customers should come from word-of-mouth and local medical society relationships. Once you have 20 practices generating consult data and reimbursement metrics, you have enough to approach a regional health plan for a pilot value-based contract.

The expansion playbook is specialty-by-specialty. Start with cardiology and endocrinology (highest volume, most routine referrals, best eConsult evidence base). Add dermatology next (image-based consults are a natural fit for AI and asynchronous review).

Then nephrology, rheumatology, and gastroenterology. Each new specialty requires recruiting specialists to the consult panel and training the AI on specialty-specific guidelines and case patterns. Each new specialty also opens up a new market segment and a new revenue stream.

One important competitive moat to build early: the consult database. Every AI-assisted consult and every specialist eConsult interaction is a training data point. Over time, the platform accumulates a proprietary dataset of real-world clinical questions, AI recommendations, specialist responses, and patient outcomes. This flywheel is the most defensible asset the company will own. Nobody else will have it because nobody else is capturing this data at this level of granularity across these clinical workflows.

Why investors should care

The market size is large and growing. Over 100 million specialist referrals per year in the US, at an average downstream cost of nearly \$1,000 each. That is a \$100 billion referral economy, a meaningful chunk of which is waste. If the platform captures 1% of that market through per-consult fees and shared savings arrangements, that's a \$1 billion revenue opportunity.

The timing is right. UnitedHealthcare's 2026 referral requirements are a massive signal that the payer world wants PCPs to be more active gatekeepers. CMS is building out reimbursement codes for interprofessional eConsults. AI clinical decision support tools are maturing rapidly, with major EHR vendors expected to release AI documentation and clinical support tools throughout 2026. The infrastructure layer is ready for someone to build this application on top of it.

The capital efficiency is attractive. This is not a hardware company or a drug company. It is a software platform with a services layer (the specialist network). The MVP can be built for under \$500K. First revenue can come within 6-9 months of launch. The specialist network is variable cost (paid per consult, not salaried). Technology costs are declining rapidly. And the data moat grows organically with use.

The competitive landscape is surprisingly thin for this specific model. RubiconMD and AristaMD have built eConsult platforms, but neither has a strong AI triage. Glass Health and OpenEvidence are building AI clinical decision support, but neither has a specialist consult network. Nobody has put the two together with a PCP-founder-led go-to-market. That gap is the opportunity.

The exit paths are clear. A company like this is an acquisition target for EHR vendors (Epic, Oracle Health), payer-services companies (Optum, Evernorth, Elevance), or primary care platforms (Oak Street, Agilon, Privia). It could also build into a standalone public company if the data flywheel and network effects prove out.

Where this breaks and what to watch

The biggest risk is AI accuracy. If the triage engine keeps cases in primary care that should have been referred, patients get hurt and the company dies. The calibration of the triage function needs to be conservative initially (over-refer rather than under-refer) and progressively tuned with real outcome data. This is not a “move fast and break things” situation. This is medicine.

The second risk is specialist network quality and retention. Specialists need to respond quickly and provide high-quality recommendations, or PCPs will stop using the platform. Compensation needs to be fair, and the workload needs to be manageable. Burnout in the specialist panel is a real operational risk.

The third risk is EHR integration. As long as the platform requires PCPs to use separate workflow outside their EHR, adoption friction will limit growth. Achieving native EHR integration (especially with Epic, which controls roughly 38% of the ambulatory market) is critical for scaling beyond early adopters. This is an expensive and time-consuming process, which is another reason to start with a standalone workflow and prove value before investing in deep integrations.

The fourth risk is payer politics. If payers decide that AI-assisted primary care should be reimbursed at lower rates (because the PCP is “just following the AI”), the economic model breaks. Engaging with payers early and framing the AI as an

enhancement to clinical judgment (not a replacement) is critical for protecting reimbursement.

And the fifth risk is the competitive response from incumbents. If Epic builds this natively into its platform, or if a large payer-services company launches a competing offering, the standalone startup has a harder path. Speed matters. Getting to 100 practices and a meaningful consult database before the big players move is the strategic imperative.

None of these risks are fatal. All of them are manageable with the right team, the right capital, and the right clinical leadership. The PCP who builds this company has an unfair advantage: they live in the workflow, they feel the pain, and they have the clinical credibility to recruit both the specialist network and the PCP customer. That is rare in health tech, where most companies are built by people who have not set foot in an exam room. The doctor-founder who can bridge the clinical and business worlds, and who is willing to surround themselves with people who know what they don't know, is exactly the kind of entrepreneur this market is waiting for.



5 Likes • 1 Restack

[← Previous](#)

Discussion about this post

Comments

Restacks



Write a comment...