PHYSICIAN EXTENDER SERVICE (PES).

The Space Between Visits

A Physician-Led, AI-Enabled Model for Clinical Intelligence, Responsible Automation, and Efficient Scale

A STRATEGIC WHITE PAPER ON THE PHYSICIAN EXTENDER SERVICE (PES)

Dr. Craig Coleman.
Founder, Coleman Health Services

Contents

Executive Summary

High-risk patients with chronic disease spend a small fraction of their lives in direct contact with a physician. In 2018, a chronic condition was the major reason for 39 percent of office-based physician visits nationally, and adults with two or more chronic conditions accounted for the majority of primary care visits as of 2015 [1]. Between those encounters, symptoms shift, medications lapse, and function declines quietly. When that deterioration becomes visible, it is often because the patient has called urgently, presented to an emergency department, or been admitted.

This is not primarily a data problem. Electronic health records, patient portals, remote patient monitoring (RPM) and remote therapeutic monitoring (RTM) programs, and care management platforms already collect substantial information about patients between visits. The unresolved problem is converting that information into prioritized, clinically meaningful intelligence, delivered to the treating clinician, without adding another dashboard, alert stream, or documentation burden to an already strained physician workforce.

The Physician Extender Service (PES), developed by Coleman Health Services (CHS), is a tiered clinical-intelligence architecture built to close that gap for selected high-risk chronic disease populations. Trained Patient Engagement Specialists conduct structured, disease-specific outreach between visits. Ambient AI captures and synthesizes each encounter and proposes an initial Green, Yellow, or Red classification.

A registered nurse reviews every encounter and validates or changes that classification using licensed clinical judgment, managing appropriate Yellow-level coordination.

RN-validated potential Red encounters receive physician review and clinical interpretation, and a concise, clinically validated summary — with a physician-to-physician recommendation when physician review is indicated — reaches the treating clinician, who retains full authority over diagnosis, prescribing, treatment, and longitudinal management.

PES is designed to complement, not duplicate or displace, the population health, care management, RPM/RTM, and physician infrastructure that health systems, ACOs, health plans, and federal healthcare programs already operate.

The paper's central argument extends beyond the clinical model itself.

PES is structured from launch to generate the evidence needed to responsibly increase the role of automation over time. Universal human review in the current operating model is not the intended permanent state; it functions simultaneously as a clinical safeguard and as a validation mechanism that produces the reference data against which AI performance can later be measured.

As that evidence accumulates, future PES deployments are designed to evolve in defined stages — from a human-led, AI-assisted model toward AI-classified, human-audited operation, and eventually, for well-defined populations and use cases, toward AI-conducted, human-supervised operation.

Each proposed increase in automation is governed by a single standing principle: AI autonomy must be earned through demonstrated performance, not assumed on the basis of technical capability. The intended progression is evidence, then confidence, then controlled autonomy, then efficient scale — and no claim is made here that this progression has already occurred.
That governance pathway is what allows PES to address a second, structural question in healthcare delivery: whether the traditional linear relationship between patient volume and workforce growth can be loosened.

Chronic disease prevalence continues to rise as the population ages, while the Association of American Medical Colleges projects a national physician shortage of between 13,500 and 86,000 physicians by 2036, including a shortfall of 20,200 to 40,400 primary care physicians [2].

A service model that requires proportional staffing growth for every increment of patient growth cannot close the between-visit intelligence gap at the scale the problem demands. Whether PES can, over time and under governance, make that relationship less linear — without asking any organization to trade oversight for scale — is the central question a pilot deployment is designed to begin answering. This paper does not claim that question is already resolved.

PES is a physician-led clinical intelligence layer — not another data feed, not a monitoring device, and not an autonomous diagnostic system. It is designed to earn expanded automation through evidence, not to assume it.

The remainder of this paper describes the clinical problem PES addresses, its current operating model and governance architecture, the staged pathway future deployments could follow toward greater automation, and the strategic implications for health systems, value-based care organizations, health plans, and federal healthcare programs. It closes by distinguishing explicitly what the current operating model does today, what a pilot is designed to implement and evaluate, and what remains future-state architecture contingent on evidence not yet gathered.

01 The Clinical Blind Spot Between Visits

Chronic disease management in the United States is organized around an episodic visit structure. National ambulatory care survey data show that adults with two or more diagnosed chronic conditions made up 53.1 percent of primary care office visits as of 2015, and a chronic condition was the major reason for 39 percent of all office-based physician visits in 2018 [1]. A separate analysis of National Ambulatory Medical Care Survey data from 1997 through 2005 found that the number of clinical items addressed during an adult primary care visit — diagnoses, medications, tests, and counseling combined — rose from 5.4 to 7.1 over that period, even as visit duration grew more slowly, from 18.0 to 20.9 minutes [3]. The volume of clinical work packed into each visit has tended to outpace the time available for it. The structural mismatch is straightforward: chronic disease requires continuous management, but the delivery system is organized around discrete, time-limited encounters separated by weeks or months.

In the interval between those encounters, a patient's clinical status is not static. Medications are missed or discontinued. Symptoms of heart failure, COPD, diabetes, or hypertension worsen gradually, in ways that are individually unremarkable but cumulatively significant. Weight shifts, functional status declines, and early warning signs accumulate without a structured mechanism to surface them to the treating clinician before they become acute. An evidence-based analysis of ambulatory care continuity and fragmentation, prepared for Health Quality Ontario, reviewed studies of patients with diabetes, heart failure, and COPD and found that studies of ambulatory care fragmentation have consistently reported greater fragmentation to be associated with greater subsequent emergency-department utilization, though the authors graded the overall quality of this evidence as low under GRADE criteria [4]. The between-visit interval is not a passive gap; the available evidence, while limited, points toward it as a period where preventable deterioration can develop unseen.

Data Collection Is Not the Same Problem as Clinical Intelligence

It would be a mistake to characterize this as primarily a data-collection problem. Health systems, ACOs, and health plans have invested heavily over the past decade in EHR infrastructure, patient portals, RPM and RTM devices, and care management staffing.

A 2025 systematic review of 40 randomized trials found low-certainty evidence that RPM may modestly reduce hospitalizations and hospital length of stay, while effects on emergency utilization were small or uncertain — the review reported a risk ratio of 0.86 for the proportion of patients hospitalized, a mean reduction of roughly 0.8 days in hospital length of stay, and little to no clear difference in outpatient or emergency visit rates, with certainty ranging from low to very low across outcomes [5]. A nonrandomized study of post-hospitalization RPM for heart failure and COPD patients within an accountable care organization similarly used nurse-provided oversight and triage, rather than device data alone, as a core component of the intervention, and reported a reduction in adjusted six-month mortality alongside a non-significant reduction in the combined outcome of death, admission, or emergency visit [6]. The pattern across this literature is consistent: monitoring technology alone produces modest and uncertain results; the studies with more structured clinical interpretation built into the intervention tend to report more consistent findings.

That pattern is the premise underlying PES. More devices, more portals, and more raw data points do not by themselves close the between-visit blind spot if nothing structured and clinically credible determines which signals matter, escalates the ones that do, and filters out the ones that do not — before any of it reaches a physician managing a growing volume of clinical items in a visit whose length has not kept pace [3].

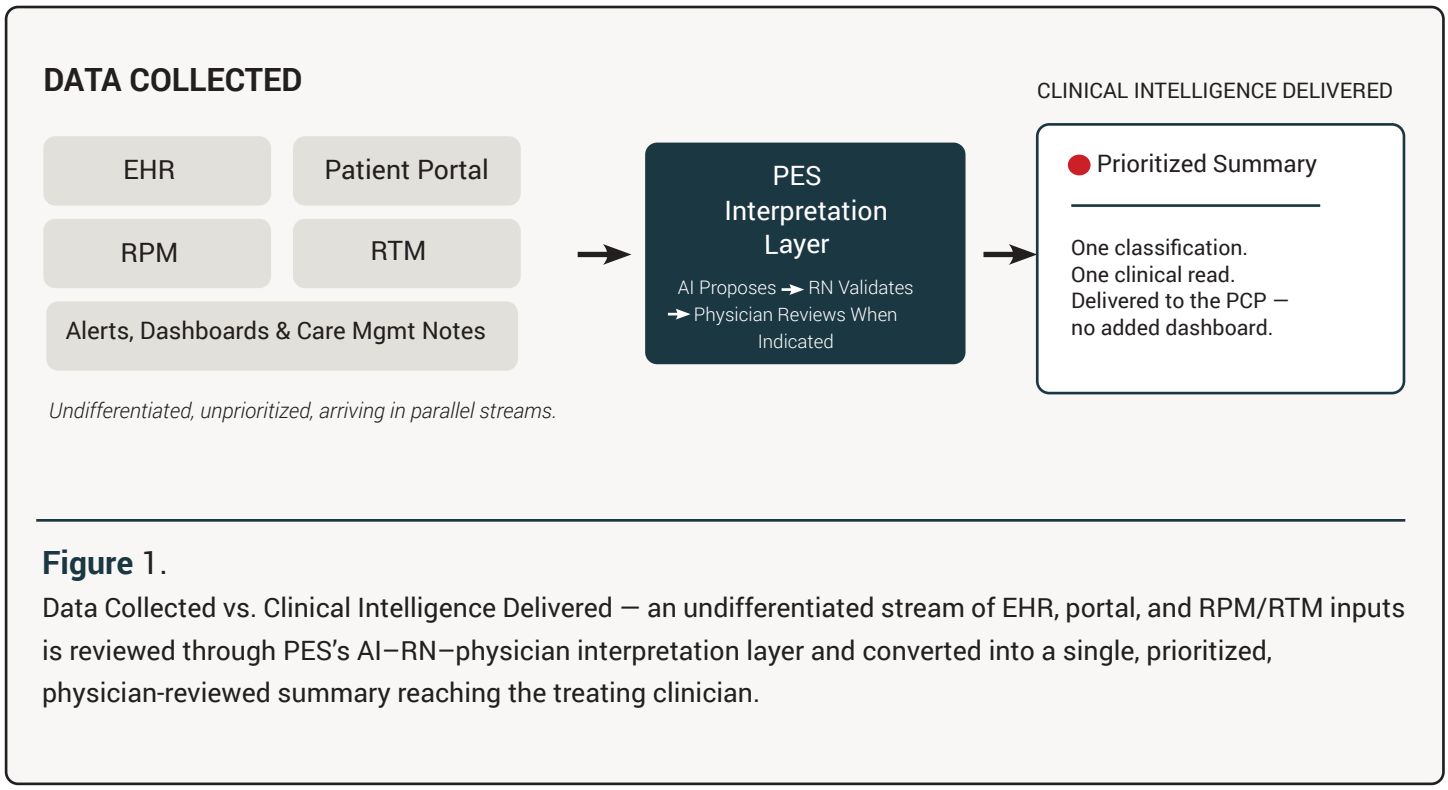


Figure 1.
Data Collected vs. Clinical Intelligence Delivered — an undifferentiated stream of EHR, portal, and RPM/RTM inputs is reviewed through PES's AI–RN–physician interpretation layer and converted into a single, prioritized, physician-reviewed summary reaching the treating clinician.

02 From Data Collection to Clinical Intelligence

PES is positioned as an interpretation layer that sits between existing data sources and the treating physician — not as a replacement for infrastructure a health system, ACO, or health plan has already built.

Population health and care management

Population health platforms stratify risk and identify cohorts; care management programs provide longitudinal coordination, education, and outreach for defined populations. PES does not replicate risk stratification or care coordination. It provides a structured, recurring clinical touchpoint for a selected high-risk cohort and channels what it learns into a form the existing care management team, and the treating physician, can act on.

RPM and RTM

Remote patient and remote therapeutic monitoring generate physiologic and device-based data that require review, context, and follow-up. PES does not compete with RPM/RTM device programs or require an organization to adopt new hardware; the current PES model begins with structured conversational outreach rather than device deployment specifically so it can be layered onto whatever monitoring infrastructure, or absence of infrastructure, an organization already has in place.

EHR infrastructure

The EHR remains the system of record. PES is designed to deliver a concise clinical summary into existing clinical workflows rather than to introduce a parallel data system competing for the PCP's attention alongside the EHR inbox.

Existing nursing resources

Health systems and ACOs already employ RNs in care management, utilization review, and triage roles. PES's RN function is a defined, protocol-driven checkpoint specific to the PES encounter — reviewing the AI-proposed summary and classification, managing appropriate Yellow-level issues, and escalating validated potential Red encounters — not a duplication of those existing roles.

The treating physician

The treating PCP remains the accountable clinical decision-maker throughout. PES does not assume responsibility for the patient's ongoing diagnosis, prescribing, treatment, or longitudinal management; those authorities remain with the treating clinician. PES reviewing physicians provide clinical interpretation and recommendations within the defined PES workflow, intended to make the PCP's next decision better-informed, not to substitute for it.

PES is intended to complement — not replace — population health, care management, RPM/RTM, EHR infrastructure, existing nursing resources, and the treating physician

03 The PES Operating Model

This section describes the current PES operating model — what the service is designed to do at launch. It reflects the most recent revision of the clinical model and supersedes earlier descriptions in which every classification decision was made directly by a physician. In the current model, AI proposes the initial classification, an RN reviews every encounter, and a physician evaluates potential Red encounters that the RN has validated and escalated.

Workflow

High-risk cohort selection — an initial focus on congestive heart failure, COPD, diabetes, and hypertension, chosen for their prevalence, their propensity for between-visit deterioration, and the relative maturity of disease-specific outreach protocols.

Structured patient outreach — a trained Patient Engagement Specialist conducts a standardized, disease-specific interview with the patient on a defined cadence.

AI capture, synthesis, and proposed classification — ambient AI documents the encounter and proposes an initial Green, Yellow, or Red classification from the structured interview content.

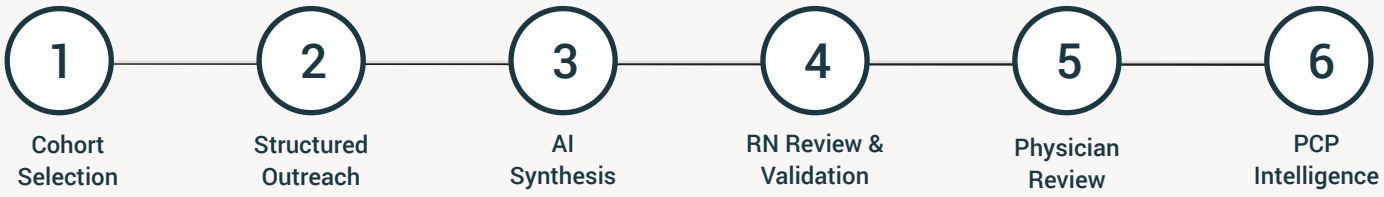
RN review and validation — a licensed RN reviews every encounter, validates or changes the AI-proposed classification using independent clinical judgment, and manages appropriate Yellow-level coordination with the patient's existing clinical team.

Physician review of RN-validated potential Red encounters — a reviewing physician independently interprets each RN-validated potential Red encounter and formulates a physician-to-physician clinical recommendation.

Communication to the treating clinician — a concise, clinically validated summary and, when physician review is indicated, a physician-to-physician recommendation, is delivered to the PCP.

Clinician authority preserved — the treating clinician retains full authority for diagnosis, prescribing, treatment, and longitudinal management at every step.

FIGURE 2



Physician-led at every step — the treating clinician retains full authority throughout. Intelligence

Figure 2.

Six-Step PES Clinical Workflow

Cohort selection → Structured outreach → AI synthesis → RN review and validation → Physician review of validated Red encounters → Communication to the treating clinician.

Green, Yellow, Red — Conceptually

AI proposes the initial classification from the structured interview content; a licensed human reviewer retains classification authority at every subsequent step.

Classification	● Green — Stable	● Yellow — Clinical Drift	● Red — Acute Risk
Meaning	No meaningful clinical change identified in the structured interview.	Potential early change: symptom shift, medication concern, or functional decline that does not clearly meet physician-review criteria.	Potential meaningful clinical deterioration.
Disposition	Routine monitoring continues; no additional review required.	RN reviews, clarifies, verifies medications, and coordinates with the patient's existing clinical team as needed.	Reviewing Physician interprets the encounter and communicates a physician-to-physician recommendation to the PCP.

Roles and Authority Boundaries

Patient Engagement Specialist — conducts the standardized interview. Does not document independently, classify, or exercise clinical judgment; role is limited to structured data capture through direct patient conversation.

Ambient AI — generates the clinical documentation and proposes the initial Green, Yellow, or Red classification from the interview content. Does not finalize a classification, communicate directly with the PCP, or take any action affecting patient care.

Registered Nurse — reviews every encounter and validates or changes the AI-proposed classification using licensed clinical judgment; manages Yellow-level coordination; escalates only validated potential Red encounters.

Reviewing Physician — independently interprets each RN-validated potential Red encounter and communicates a physician-to-physician recommendation. Does not routinely re-review Green or RN-managed Yellow encounters as part of the current operating model, though a pilot is intended to evaluate whether targeted quality-assurance sampling of those encounters adds value.

Treating PCP — retains full and exclusive authority over diagnosis, treatment, prescribing, and overall patient management. PES supplies clinical intelligence; it does not supply decisions.

Emergency Protocol

The PES model includes a defined emergency protocol: on recognition of a predefined emergency trigger during the structured interview, the Engagement Specialist stops the interview and delivers a standardized instruction directing the patient to seek emergency care. This action does not require the Engagement Specialist to make an independent clinical determination; it is a protocol-directed response to predefined triggers identified during the structured encounter, not a clinical judgment exercised by non-clinical staff. No AI or RN classification step is interposed before that instruction is given.

04 Human Oversight as Validation Architecture

The most consequential design choice in PES is not the Green/Yellow/Red schema itself, but the reason universal human review sits at the center of the current operating model. Human review in PES is designed to serve two purposes simultaneously, and both are essential to understanding why the current, labor-intensive configuration should not be mistaken for the intended permanent operating state.

Clinical safeguard

During introduction of any AI-assisted clinical workflow, comprehensive human review protects patients while the system's real-world performance is still being established. Every AI-proposed classification in the current PES model is reviewed by a licensed RN before it results in any clinical action,
and every RN-validated potential Red encounter receives independent physician interpretation before a recommendation reaches the PCP. No patient-facing clinical determination is made by AI alone.

Validation mechanism

The same universal review is also intended to generate something PES needs independent of its safety function: a growing body of human-reference judgments against which AI-proposed classifications can be systematically compared. Each RN review and each physician review is designed to produce a data point — concordant or discordant with the AI's proposed output — that a pilot is intended to use to measure the AI's actual classification performance over time, by condition, by patient subgroup, and by the type of clinical signal involved. This concordance analysis has not yet been performed for the current version of PES; it is a pilot objective, described further in Sections 9 and 10.
This dual function matters because it resolves an apparent tension in the model. A skeptical reader might reasonably ask why an organization would deploy AI at all if every output requires human review. The answer is that the initial labor structure is not the product — it is the mechanism by which
a future, more automated form of the product could be validated. Without a substantial period of universal human review, there would be no defensible basis for later reducing that review, because there would be no evidence describing how the AI actually performs against clinical reality, only evidence describing how it performs against its own training assumptions.

The current, human-intensive configuration of PES is not the intended permanent operating model. It is the validation architecture that a more automated, evidence-justified operating model would need to be built on.

This framing also carries a caution the governance architecture in Section 5 is designed to address directly. A 2017 systematic review of automation bias in clinical decision support found that the tendency of human reviewers to over-rely on automated outputs — accepting them without independent verification — is a documented risk across clinical decision support tools generally [7], and an earlier review specific to clinical decision support systems described the same pattern [8]. A small experimental study of AI-assisted tumor-cell estimation in computational pathology similarly found that the presence of an AI recommendation could shift experienced reviewers toward incorrect judgments in a measurable share of cases [9]. A validation architecture that assumes RN and physician reviewers are functioning as fully independent checks cannot take that independence for granted; it needs to be designed, measured, and monitored for automation bias as an active operating risk, not an incidental one.

05 Responsible AI Governance by Design

PES treats AI governance as a structural component of the clinical model, designed in from the outset, rather than a compliance layer added after deployment. The paragraphs below describe the governance architecture PES is designed to operate under; where a mechanism is planned for pilot implementation rather than already in continuous production use, that distinction is noted explicitly.

Defined AI authority boundaries

The AI's role is explicitly bounded: it documents the structured interview and proposes an initial classification. It is not designed to have authority to finalize a classification, to communicate directly with a PCP, or to take any action affecting patient care without an intervening licensed human decision-maker. The operating model calls for every boundary to be auditable — traceable to a specific AI output, a specific human review or change, and a specific accountable individual.

AI/human concordance monitoring

Because every AI-proposed classification is reviewed by an RN, and every RN-validated potential Red encounter is reviewed by a physician, PES is designed to track concordance between the AI's proposed classification and the human reviewer's final determination — by disease category, by classification tier, and over time. The pilot described in Sections 9 and 10 is intended to be the first structured collection of this concordance data for the current, AI-assisted version of PES; it does not yet exist for this version of the model. Discordance is intended to be treated as a governance signal, not merely a quality metric to minimize — understanding where and why the AI and human reviewers disagree is central to any future decision about where automation could safely expand.

False positives, false negatives, and asymmetric concern

Not all classification errors carry equal clinical weight. A false positive — an AI-proposed Yellow or Red that a human reviewer downgrades — costs reviewer time but does not directly endanger a patient. A false negative — an encounter the AI proposes as Green that in fact represented meaningful clinical drift — is the error category of greatest concern, because it is the one most likely to allow the between-visit blind spot the model exists to close. The governance architecture calls for false-negative detection and review to be weighted more heavily than false-positive management, consistent with the relative clinical stakes; the pilot is intended to generate the data needed to evaluate whether this weighting is being achieved in practice.

RN review, physician disposition, and exception management

Every RN review and change to an AI-proposed classification, and every physician disposition of an RN-escalated potential Red encounter, is intended to be captured as a discrete, reviewable event. Exception cases — ambiguous presentations, conflicting information, technical or communication failures during the structured interview — are designed to route to defined escalation paths rather than being resolved silently within the AI layer. The emergency protocol described in Section 3 illustrates this design principle directly: on a predefined trigger, the Engagement Specialist stops the structured interview and delivers a standardized instruction to seek emergency care, with no AI or RN classification step interposed.

Performance monitoring, drift detection, and change management

AI systems are not static once deployed; underlying model updates, prompt changes, workflow revisions, or shifts in the patient population can all change how a system performs. The PES governance architecture calls for ongoing monitoring of classification concordance and error patterns, formal change management whenever the model, prompts, or workflow are materially modified, and revalidation of performance following any such change rather than an assumption that prior validation still holds. Automated drift-detection tooling and continuous production monitoring infrastructure are part of this designed architecture; they are not represented here as already operating at scale, and their build-out is expected to track the pilot and early deployment period.

Auditability and version control

The operating model calls for every AI-proposed summary and classification, every human review and change, and every version of the underlying model, prompts, and workflow configuration to be logged and traceable, so that any individual encounter's decision history — and the system version that produced it — can be reconstructed after the fact. Building this logging and version-control infrastructure to production maturity is part of the pilot and early deployment work, not a claim that a mature audit trail already exists.

Alignment with recognized frameworks

This architecture is designed with reference to the structure of the NIST AI Risk Management Framework, which organizes AI risk management around four functions — Govern, Map, Measure, and Manage — with Govern operating as a cross-cutting foundation and Map, Measure, and Manage applied iteratively across an AI system's lifecycle [10]. It is likewise designed to align structurally with ISO/IEC 42001, an organizational AI management-system standard addressing risk management, accountability, lifecycle governance, and continual improvement for organizations that develop, provide, or use AI systems [11]. CHS founder Dr. Craig Coleman holds ISO/IEC 42001 Lead Auditor certification. PES is not represented here as formally certified against ISO/IEC 42001 or as making a compliance claim under the NIST AI RMF; both are described as frameworks the architecture is designed to align with in structure and intent.

06 Automation Must Be Earned

PES is organized around a single governing principle for expanding the AI's role over time: technological capability alone does not authorize autonomous operation. An AI system's technical capacity to perform a task is a necessary but not sufficient condition for allowing it to perform that task with reduced human oversight.

Each proposed increase in AI autonomy within PES is designed to be evaluated against four categories of evidence:

Performance

demonstrated concordance with human reviewers across a sufficient volume of encounters, disease categories, and patient subgroups to support a defensible conclusion about accuracy, not an anecdotal one.

Clinical risk

the severity and reversibility of harm a classification error in the relevant category could plausibly cause, with particular weight on the false-negative risk described in Section 5.

Monitoring capability

whether functioning mechanisms exist to detect degraded performance after the change, not only before it.

Recovery and reversibility

whether an automation increase can be rolled back cleanly if post-implementation monitoring reveals a problem, without harm accruing in the interim.

The intended progression is evidence, then confidence, then controlled autonomy, then efficient scale — in that order, without a step skipped because a deadline, a sales cycle, or a competitive pressure makes an earlier stage inconvenient.

This principle distinguishes the staged automation roadmap described in the following section from a simple plan to eventually replace human reviewers with AI. The roadmap describes conditions under which automation could expand in future deployments; it does not guarantee those conditions will be met on any particular timeline, and it does not treat automation of one clinical function as evidence supporting automation of a different one.

Automation of one function does not constitute validation of another.

07 Technology Evolution: Validate. Optimize. Scale.

Future PES deployments are designed to be capable of evolving through three conceptual stages, each unlocked only by evidence generated in the stage before it. No specific timeline is attached to these stages; the transition between them is intended to be governed by the evidentiary conditions in Section 6, not by a fixed schedule. Only Stage 1 describes the current operating model; Stages 2 and 3 describe future-state architecture that would require its own validation before implementation.

Stage	Operating Model	Governing Question
Stage 1 — Human-Led, AI-Assisted (current model)	Engagement Specialist conducts outreach; AI captures, summarizes, and proposes classification; RN reviews and validates every encounter; physician reviews RN-validated potential Red encounters.	Validate: does the AI's proposed classification hold up against universal human review?
Stage 2 — AI-Classified, Human-Audited (future state)	Only after sufficient validated concordance is demonstrated, RN oversight could shift from universal review of every encounter toward statistical sampling, low-confidence review, high-risk review, discordance review, and targeted exception audits	Optimize: can equivalent safety be maintained with a more targeted, risk-weighted review model?
Stage 3 — AI-Conducted, Human-Supervised (future state)	Following separate validation specific to this distinct function, conversational AI or AI agents could conduct certain structured interactions directly for defined patient populations and use cases, with human handoff for ambiguity, low confidence, patient preference, communication difficulty, exceptions, or any circumstance outside the validated operating envelope.	Scale: can defined, lower-complexity interactions be conducted by AI without degrading safety or patient experience?

Two aspects of this progression deserve emphasis. First, Stage 2 would not eliminate human review — it would reallocate it toward the encounters most likely to need it, using concordance and discordance data accumulated in Stage 1 to determine what "most likely to need it" actually means for a given condition or population, rather than assuming it in advance. Second, Stage 3 concerns a different function than Stages 1 and 2: automating the conduct of the structured interview itself is a distinct capability from automating its classification, and under the principle established in Section 6, validating one would not validate the other. Movement to Stage 3 for any given population or use case would require its own evidentiary basis, evaluated independently, and is not part of the current pilot design.

Figure 3.
Three-Stage Technology Evolution Roadmap. Each stage is unlocked only by evidence generated in the stage before it; no timeline is implied, and later-stage automation is not represented as currently deployed or proven.

08 Efficient Scale and the Labor Curve

Traditional healthcare service models tend to scale along a largely linear curve: more patients require more encounters, which require more personnel, which require more cost, in roughly fixed proportion. That relationship constrains how much between-visit clinical intelligence any organization can realistically deliver, particularly against the physician supply picture described in the Executive Summary — a national shortage projected at up to 86,000 physicians by 2036, concentrated most heavily in primary care and in rural and underserved areas already designated as health professional shortage areas [2, 12].

The staged automation roadmap in Section 7 is the mechanism by which future PES deployments — strictly under the governance conditions described in Sections 5 and 6 — could work toward separating the growth of the patient population served from the growth of the human workforce required to serve it. This is proposed as workforce reallocation toward the work that most requires human judgment, not as workforce reduction, and it remains a hypothesis to be tested rather than a demonstrated result.

- Engagement Specialists could increasingly concentrate on exception handling, accessibility accommodations, and patients who need or prefer direct human interaction, rather than conducting every structured interview personally as volume grows.
- RNs could increasingly perform targeted coordination, quality assurance, and exception management, applying nursing judgment where evidence indicates it adds the most value, rather than reviewing every encounter regardless of risk.
- Reviewing physicians could concentrate their attention on cases that genuinely require physician-level interpretation, rather than serving as a routine first pass across a growing volume of low-acuity encounters.
- Treating PCPs would continue to receive a single prioritized stream of clinical intelligence rather than an additional, undifferentiated data feed layered on top of the EHR inbox they already manage.

Under this framing, the appropriate long-term measure of PES's value is not simply the number of patients it monitors. It is a different, more demanding question: how many patients could PES safely support without requiring its human infrastructure to grow at the same rate?

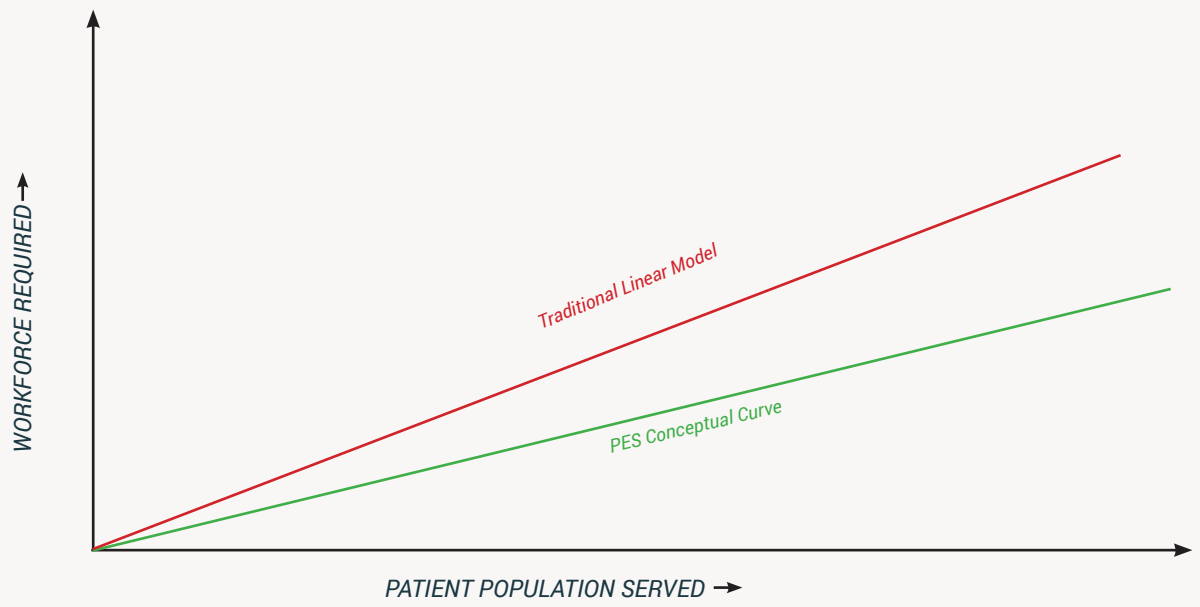


Figure 4.

Traditional Linear Labor Curve vs. PES Conceptual Curve. A conceptual illustration only — not a financial or staffing projection, and not a demonstrated result. It contrasts a traditional labor model that scales roughly with population against a PES hypothesis in which validated automation may allow capacity to grow faster than headcount

That is a question the pilot described in Sections 9 and 10 is designed to begin answering — not a claim this paper makes in advance.

09 The Pilot: Clinical Evaluation and Learning System

The proposed entry point for organizations engaging with PES is a defined 6- to 12-month pilot with a selected high-risk cohort. The pilot is intended to serve two purposes at once: a clinical and operational evaluation of the PES service itself, and a learning system that generates the human-reference validation data described in Section 4.

What the pilot is designed to evaluate

- **Workflow feasibility** — whether structured outreach, AI synthesis, RN review, and physician review function as intended in a live operational setting.
- **Operational integration** — how PES fits into the host organization's existing clinical and administrative workflows, with no requirement to change existing systems.
- **Patient participation and experience** — enrollment, retention, and patient experience with the structured outreach model.
- **Treating-clinician usefulness** — whether PCPs find the delivered clinical intelligence timely, actionable, and appropriately concise.
- **Green, Yellow, and Red distribution** — the rate and pattern of classifications, and how that distribution compares across disease categories.
- **RN and physician workload** — the actual time burden of review, and whether the current staffing configuration is sustainable at projected scale.
- **AI/human concordance and false-negative and false-positive patterns** — the validation data central to the governance architecture in Sections 4 and 5.
- **Safety indicators** — near-miss and adverse-event tracking specific to the PES workflow.
- **Utilization outcomes, where appropriately measurable** — ED visits, hospitalizations, and related metrics, understood as directional signals from a pilot-scale, non-randomized cohort rather than controlled-trial evidence.
- **Economic scalability** — early, directional indications relevant to the labor-curve question raised in Section 8.

Distinguishing prior experience from what the pilot must establish

It is important to distinguish what CHS's prior operational experience supports from what remains to be established. The clinical operating model described in Section 3 draws on early experience with a physician-led chronic disease outreach model at a U.S. Navy clinic in Bahrain in 2016, involving approximately 50 patients with hypertension and type 2 diabetes. That experience is properly characterized as informal operational experience that informed development of the current PES model — not as a controlled clinical trial, and not as a source of validated outcome statistics for the current AI-assisted version of PES, which did not exist in that form at the time. What a defined pilot is intended to establish — and what the Bahrain experience alone cannot establish — is how the current, AI-assisted, RN-reviewed PES model performs against defined clinical, operational, and economic metrics in a modern deployment.

10 What the Pilot Is Designed to Establish

The propositions below are the specific evidence a pilot deployment is intended to generate. They are stated here as open questions PES is designed to answer, not as findings already established.

- Whether AI-proposed initial classification in the current PES workflow achieves concordance with RN and physician review at a level sufficient to support progression toward Stage 2 (AI-Classified, Human-Audited). This concordance has not yet been measured for the current model version.
- Whether the current allocation of Engagement Specialist, RN, and physician workload is sustainable at scale. This is an operational hypothesis, not a validated staffing ratio.
- Whether RN and physician reviewers in the PES workflow are subject to material automation bias in practice, and what monitoring or workflow design most effectively limits it. The general literature on automation bias in clinical decision support indicates this is a real risk requiring active measurement [7, 8, 9], not an assumption resolved by design alone.

- Whether PES is associated with measurable differences in ED utilization, unplanned admissions, or related utilization outcomes. The external RPM literature cited in Section 1 shows directional, low-to-very-low-certainty evidence for monitoring-plus-triage models generally [5, 6]; it does not establish PES-specific outcomes, which remain a pilot hypothesis.
- Whether the economic and workforce-scaling objective described in Section 8 — breaking the linear relationship between patient growth and workforce growth — is achievable, and at what rate or magnitude. No financial projections, ROI estimates, or staffing-ratio commitments should be inferred from this paper pending that evidence.
- What timeline, if any, is realistic for progression through the stages described in Section 7. Progression is evidence-gated by design; no specific timeline is represented here or should be represented to prospective partners in advance of pilot results.
- The Bahrain Navy Clinic experience described in Section 9 should not be read as prior evidence toward any of the propositions above; it predates the current AI-assisted model and is presented in this paper strictly as background on how PES was developed, not as validation data.

11 Strategic Applications

The architecture described in this paper — a physician-led interpretation layer with a governed pathway toward greater automation — has different but overlapping relevance across several categories of healthcare organization.

Health systems and physician enterprises

For organizations managing physician capacity as a scarce, expensive resource, PES is designed to extend between-visit visibility while concentrating physician capacity on cases requiring physician-level judgment, keeping physician judgment as the final word on any Red-tier encounter.

ACOs and value-based care organizations

Organizations bearing financial risk for avoidable utilization have a direct interest in earlier detection of between-visit deterioration, consistent with the low-quality but consistent evidence linking greater care fragmentation to greater subsequent ED utilization described in Section 1 [4]. PES's tiered model is designed to support that objective without requiring new monitoring infrastructure.

Health plans and risk-bearing organizations

For payers and other risk-bearing entities, a physician-interpreted intelligence layer offers a complement to existing utilization management and care management programs, oriented toward earlier, lower-acuity intervention rather than downstream utilization review.

Rural health, RHCs, and FQHCs

Rural counties account for a disproportionate share of the nation's primary care health professional shortage areas: a November 2025 Commonwealth Fund analysis found that 92 percent of rural counties were designated primary care health professional shortage areas in 2023, compared with 83 percent of nonrural counties, affecting an estimated 42 million residents [12]. Peer-reviewed research has found that Federally Qualified Health Centers measurably reduce the primary care provider gap in shortage counties [13]. For these settings, a model that extends physician-level interpretation without requiring an on-site physician for every touchpoint addresses a structural access constraint directly.

Federal healthcare systems — VA, DoD, IHS

Federal healthcare systems face many of the same structural pressures — geographically distributed patients, constrained physician staffing, and a strong policy interest in avoidable-utilization reduction — that shaped the operational precedent for PES. CHS's founder brings direct federal healthcare operations experience relevant to deploying an AI-assisted clinical system within federal regulatory and security requirements.

Organizations with existing RPM, care management, or population-health infrastructure

For organizations that have already invested in these capabilities, PES's differentiator is a physician-led interpretation layer designed to work with what already exists, rather than a new system that would require replacing it.

12 Conclusion

Healthcare has become reasonably effective at collecting data about patients between visits. It has not yet become reliably effective at converting that data into clinical action before the next scheduled appointment becomes an emergency department visit. That is the specific, bounded problem PES is built to address.

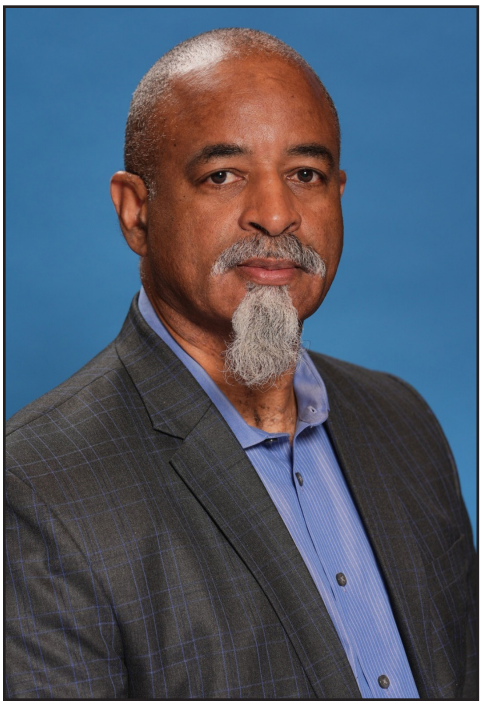
The deeper argument of this paper concerns how technology should be allowed to take on more of that work over time. Technology should not merely digitize an already inefficient human workflow. Its purpose, responsibly pursued, is to progressively absorb routine, repeatable work — after, and only after, it has demonstrated the reliability to do so — so that scarce physician and nursing expertise can migrate toward the ambiguous cases, the exceptions, the relationships, and the judgment calls that genuinely require it.

PES is offered as one credible architecture for pursuing that goal without asking any health system, payer, or federal healthcare program to trade clinical accountability for efficiency. The relationship it is built around is straightforward:

Responsible Governance + Clinical Judgment + Validated Automation = Efficient Scale

Coleman Health Services is available to discuss pilot design, partnership structures, and how PES could be tailored to a specific organization's existing infrastructure and patient population.

About the Author



Dr. Craig Coleman is the founder of Coleman Health Services and the architect of the Physician Extender Service. A physician, systems engineer, and healthcare technology leader, Dr. Coleman brings more than three decades of military, clinical, technical, and operational experience to the challenge of improving healthcare delivery.

A 30-year U.S. Navy veteran, his career has spanned engineering and complex systems, operational medicine, and federal healthcare. His work today focuses on the intersection of clinical care, artificial intelligence, and responsible technology adoption. He holds ISO/IEC 42001 Lead Auditor certification in AI management systems and has advanced training in artificial intelligence in healthcare.

His work on PES reflects a central belief: technology should extend clinical judgment, not replace it—and greater automation should be earned through evidence.

References

1. Ashman JJ, Santo L, Okeyode T. Characteristics of Office-Based Physician Visits, 2018. NCHS Data Brief No. 408. Hyattsville, MD: National Center for Health Statistics; 2021. DOI: 10.15620/cdc:105509.
2. GlobalData Plc. The Complexities of Physician Supply and Demand: Projections From 2021 to 2036. Prepared for the Association of American Medical Colleges. Washington, DC: AAMC; March 2024.
3. Abbo ED, Zhang Q, Zelder M, Huang ES. The Increasing Number of Clinical Items Addressed During the Time of Adult Primary Care Visits. J Gen Intern Med. 2008;23(12):2058–2065. DOI: 10.1007/s11606-008-0805-8.
4. Health Quality Ontario. Continuity of Care to Optimize Chronic Disease Management in the Community Setting: An Evidence-Based Analysis. Ont Health Technol Assess Ser. 2013;13(6):1–41. PMID: 24167540.

5. Smedslund G, Østerås N, Hestevik CH. Effects of Remote Patient Monitoring on Health Care Utilization in Patients With Noncommunicable Diseases: Systematic Review and Meta-Analysis. JMIR Mhealth Uhealth. 2025;13:e68464. DOI: 10.2196/68464. PMID: 41032865.
6. Harris S, Paynter K, Guinn M, Fox J, Moore N, Maddox TM, Lyons PG. Post-Hospitalization Remote Monitoring for Patients With Heart Failure or Chronic Obstructive Pulmonary Disease in an Accountable Care Organization. BMC Health Serv Res. 2024;24:69. DOI: 10.1186/s12913-023-10496-6. PMID: 38218820.
7. Lyell D, Coiera E. Automation Bias and Verification Complexity: A Systematic Review. J Am Med Inform Assoc. 2017;24(2):423–431. DOI: 10.1093/jamia/ocw105. PMID: 27516495.
8. Goddard K, Roudsari A, Wyatt JC. Automation Bias – A Hidden Issue for Clinical Decision Support System Use. Stud Health Technol Inform. 2011;164:17–22. PMID: 21335682.
9. Rosbach E, Ganz J, Ammeling J, Riener A, Aubreville M. Automation Bias in AI-Assisted Medical Decision-Making Under Time Pressure in Computational Pathology. In: Palm C, et al., eds. Bildverarbeitung für die Medizin 2025 (BVM 2025). Informatik aktuell. Wiesbaden: Springer Vieweg; 2025. DOI: 10.1007/978-3-658-47422-5_27.
10. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: U.S. Department of Commerce; January 2023. DOI: 10.6028/NIST.AI.100-1.
11. International Organization for Standardization / International Electrotechnical Commission. ISO/IEC 42001:2023 — Information Technology — Artificial Intelligence — Management System. Geneva: ISO; 2023.
12. Horstman C, Shah A. The State of Rural Primary Care in the United States. The Commonwealth Fund. November 17, 2025.
13. Xue Y, Greener E, Kannan V, Smith JA, Brewer C, Spetz J. Federally Qualified Health Centers Reduce the Primary Care Provider Gap in Health Professional Shortage Counties. Nurs Outlook. 2018;66(3):263–272. DOI: 10.1016/j.outlook.2018.02.003. PMID: 29685322.

This paper also draws on internal CHS source materials for the Physician Extender Service, including the PES Executive Summary and PES Capability Statement (2026), and earlier CHS PES Executive Brief and Executive Concept Brief materials used for strategic background. These internal materials are not independently citable and are not included in the numbered list above.