medical sciences

MDPI

Article

Artificial Epidemiology: A Decision-Theoretic Framework for Measuring AI-Induced Prevalence Divergence in Administrative Health Data

Florian O. Stummer [1,*], Hans Thulesius [1,2,3] and Thomas Frese [1]

[1] Institute of General Practice and Family Medicine, Medical Faculty, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, 06112 Halle (Saale), Germany
[2] Department of Clinical Sciences Malmö, Family Medicine, Lund University, 202 13 Malmö, Sweden
[3] Research and Development Centre, Region Kronoberg, 351 88 Växjö, Sweden
* Correspondence: florian.stummer@uk-halle.de

Abstract

Background/Objectives:AI-assisted documentation and coding tools add diagnoses to health records, and self-updating AI agents may soon learn from the codes they generate. The objective was to develop a framework for asking whether such systems could distort disease prevalence at population scale, and a measurement approach for detecting it. **Methods:** We specify a partially observable Markov decision process in which a coding agent observes the administrative record exactly but clinical state only with delay and noise, derives the conditions under which policy-gradient learning follows the administrative reward, and states the general claim as a falsifiable hypothesis. We define the artificial epidemiology divergence D(C) = P_a(C) − P_c(C) relative to an explicit clinical reference standard, relate it to sensitivity and specificity of administrative case definitions, and propose five markers with falsification conditions. A pre-study on nine synthetic datasets (Synthea; four US states; eight distinct populations, one of 11,475 and seven of about 3400 patients, and one repeated generation with the same seed) measured D(C) for five metabolic conditions without any AI coding process, with bootstrap intervals and sensitivity analyses, and tested whether injected coding changes were detectable. **Results:** Under the primary reference definition, D(C) was +0.041 (95% CI 0.036 to 0.045) for type 2 diabetes, −0.125 for hypertension, −0.191 for hyperlipidaemia and +0.314 for obesity, with consistent signs across all populations. The magnitude, and for hypertension, hyperlipidaemia and obesity the sign, depended on the reference definition. An injected upcoding of 5% per year was detected as a change in the slope of D(C) in all populations, and an undercoding of 10% per year in all populations. **Conclusions:** The baseline value of D(C) reflects the measurement definition, and the operationally relevant quantity is its change under a fixed definition. The synthetic results establish a measurement approach and its detection limits, not evidence that AI causes prevalence distortion; computing D(C) on linked administrative and clinical records around AI deployment is the necessary next step.

Keywords: artificial epidemiology; clinical coding; AI-assisted documentation; disease prevalence; partially observable Markov decision process; misclassification; electronic health records; health data governance

Check for updates

Academic Editor: Antonio Artigas

Received: 14 October 2026
Revised: 2 November 2026
Accepted: 9 November 2026
Published: 12 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In a retrospective cohort of 20,302 primary care notes, Castro et al. found that AI-scribed visits produced greater neuropsychiatric symptom documentation but a lower likelihood of documented psychiatric interventions (adjusted OR 0.83, 95% CI 0.72 to 0.95) [1]. More documentation did not produce more clinical action. It produced more administratively structured documentation. This pattern, in which AI amplifies the administrative surface of clinical data while the clinical substance remains unchanged or diminishes, is the starting point for the present paper.

The pattern now has a price. In September 2026 the Blue Cross Blue Shield Association reported that hospitals had classified 55,158 additional inpatient cases as complex between the first quarter of 2023 and the last quarter of 2025, adding about US$942 million in claims costs over two years; 70% of the increase came from secondary diagnoses that move a claim into a higher-paying diagnosis-related group without a corresponding increase in clinical care [2]. Many of these diagnoses, such as hyponatraemia, unspecified acidosis and acute posthaemorrhagic anaemia, can be derived from a single laboratory value, and the association links their growth to AI-assisted revenue-cycle software that scans records and laboratory data for codable findings [2]. Its central test is the one formalised in this paper as the documentation-action gap: in the hospitals with the fastest coding growth, posthaemorrhagic anaemia was coded in 13.7% of major bowel procedures against 9.9% elsewhere, yet only 16.9% of the patients so coded received a transfusion, against 19.3% at peer hospitals [2]. The American Hospital Association rejected the inference, noting that the analysis did not review the underlying medical records, did not identify which claims involved AI tools, and did not fully account for changing morbidity [3]. Both positions rest on claims data, and neither can settle the question with claims data alone. The quantity in dispute is the gap between what is coded and what is clinically present, which can only be measured where coded data are linked to clinical evidence. An earlier version of the present analysis was posted as a preprint in April 2026 [4].

The reification feedback loop [5] describes a self-reinforcing cycle in which documentation requirements generate coded data, coded data inform policy and AI training, and policy creates new documentation requirements. Under human conditions, this loop operates at human speed: its velocity is limited by the pace of clinical documentation, the attentiveness of coding professionals, and the frequency of regulatory revision. The distortions are real (Crespin et al. attributed up to two-thirds of the growth in highest-intensity US hospital discharges in five states to upcoding [6]), but they accumulate incrementally and remain, in principle, observable by external audit.

Self-evolving AI agents remove these speed limits. A new class of AI systems does not merely learn from data but autonomously modifies its own architectures, memory structures, and reasoning strategies in response to feedback [7]. In healthcare, such systems are being prototyped for autonomous research workflows [8], biomedical reasoning [9], and automated coding [10,11]. If such systems are deployed on data shaped by clinical–administrative negotiation, the system's outputs could become the training signal for its own next iteration, and the feedback loop could close.

The objective of this study was to develop a framework that specifies when a self-updating coding agent would follow administrative rather than clinical reward, to define a measure of the resulting divergence between administrative and clinical prevalence, and to determine, in synthetic populations without any AI coding process, the baseline of this measure and the size of coding change it can detect. The study makes three contributions: (1) a fully specified decision-process model that exposes the asymmetry between observable administrative rewards and hidden clinical outcomes, together with the conditions under which learning follows the administrative reward, stated as a falsifiable hypothesis

(Section 3.2); (2) the artificial epidemiology divergence D(C), defined relative to an explicit reference standard and related to established measures of coding accuracy and coding intensity (Section 3.3); and (3) five experimental markers with a minimal empirical programme, together with a pre-study on synthetic populations that measures the markers in the absence of any AI coding process and determines how large an injected coding change must be before it is detectable (Section 3.6, Section 3.7 and Section 4). Section 2 reviews related work and states what is new in the framework, and Section 5 discusses what the synthetic results can and cannot show, governance implications, and the conditions under which the framework would be wrong.

2. Background and Related Work

2.1. From Bias Detection to Bias Amplification

The recognition that health algorithms encode structural bias is well established. Obermeyer et al. showed that a commercial health risk prediction algorithm used healthcare costs as a proxy for health needs, producing racial bias that would have underreferred 28.8 percentage points of Black patients relative to a clinically grounded model [12]. Chin et al. subsequently argued that bias enters at the problem formulation stage, not merely through biased training data [13]. These contributions detect and characterise bias in existing, static systems. Whether self-updating systems, which modify their behaviour in response to feedback, might amplify such bias through recursive training on biased outputs is the question this paper addresses.

Feedback loops that amplify initial biases are well documented in other algorithmic domains. Recommender systems concentrate user engagement on narrowing content clusters [14], and predictive policing models direct enforcement toward historically oversurveilled areas, generating confirming arrests that reinforce the model's predictions [15]. This literature establishes the general mechanism but does not address the healthcare-specific features that make clinical-coding feedback loops distinctive: the availability of independent clinical evidence as an external check, the regulatory structure of the European Health Data Space, and the partial observability of clinical outcomes formalised in Section 3.2.

2.2. When the Classification Becomes Ground Truth

Medical terminologies, from statistical classifications such as ICD to the description-logic terminology SNOMED CT, do not merely organise health data but shape the systems trained on it. Fries et al. showed that ontology-driven weak supervision uses ICD and SNOMED CT categories to generate training labels for machine learning, elevating coding artefacts to the status of ground truth [16]. Bossen and Pine found, ethnographically, that clinical documentation integrity specialists modified physician documentation to align with coding categories rather than clinical narratives [17]. Shumailov et al. demonstrated the downstream consequence: models trained recursively on their own outputs undergo "model collapse", with rare events vanishing from the learned distribution [18]. In healthcare terms, under-coded conditions could progressively disappear from an AI system's learned world while well-coded conditions concentrate. None of these studies addresses systems that modify their own relationship to the classification.

2.3. Self-Evolving AI Agents

Gao et al. surveyed self-evolving agents along three dimensions: what evolves (models, memory, tools, architecture), when evolution occurs (within or between inference episodes), and how the system evaluates its own modifications [7]. Two healthcare-adjacent prototypes illustrate the trajectory. HealthFlow learns how to strategise about complex healthcare

research tasks, distilling procedural successes and failures into a persistent knowledge base [8]. STELLA develops an evolving library of reasoning templates for biomedical research that improve with experience [9]. Neither system performs clinical coding, but both demonstrate the architectural capability (level c in Section 3.1) that would, if applied to coding workflows, close the feedback loop.

2.4. Coding in General Practice

In daily general practice, the coding act is rarely a deliberate clinical decision. It is a byproduct of documentation under time pressure, shaped by whichever autocomplete suggestion the practice management system presents first, by the billing requirements of the quarterly settlement, and by the coding habits inherited from predecessors in the practice. In multi-physician centres (in Germany, the Medizinisches Versorgungszentrum, MVZ), this effect compounds: coding conventions become organisational culture, transmitted through shared templates and implicit norms rather than through explicit clinical reasoning. When an AI system learns from these codes, it learns the sediment of many small negotiations between clinical intent and administrative convenience, not clinical reality itself.

2.5. Axis of Novelty

A terminological clarification is needed at the outset. Tempest-Walters uses the phrase "AI epidemiology" to describe the application of epidemiological surveillance methods to AI outputs [19]. Our usage is the inverse: "artificial epidemiology" designates the epidemiological artefacts that AI operating on administratively distorted health data could create. The two concepts are complementary, not competing.

Two established lines of research measure the relation between codes and clinical reality. Validation studies estimate the sensitivity and positive predictive value of administrative case definitions against chart review at one point in time [20,21], and coding-intensity studies in health economics measure changes in coding relative to expected morbidity, typically without a clinical reference [6,22,23]. Obermeyer et al. detected bias arising from proxy selection in a static algorithm [12]; Fries et al. showed that coded categories become ground truth for training labels [16]; Shumailov et al. demonstrated recursive degradation in silico [18]; and Gao et al. surveyed self-evolution as a general capability [7]. To our knowledge, no prior work connects self-updating AI to the reification feedback loop through a decision-process model and a divergence measure designed for repeated measurement around deployment events. Table 1 summarises the comparison.

Table 1. Comparison of the present framework with the closest prior approaches.

Dimension	Code Validation [20,21]	Coding Intensity [22,23]	Obermeyer et al. [12]	Fries et al. [16]	Shumailov et al. [18]	This Paper
Measures codes against clinical evidence	Yes, at one time	No	Partial	No	No	Yes, repeatedly
Measures change in coding over time	No	Yes	No	No	No	Yes
Models classification-to-training pipeline	No	No	Partial	Yes	No	Yes
Demonstrates recursive degradation	No	No	No	No	Yes	No
Models a learning coding agent	No	No	No	No	No	Yes (formal model)
Empirically validated on real data	Yes	Yes	Yes	Yes	Yes	No

The individual components of the framework are established: partially observable decision processes, policy-gradient learning, sensitivity and specificity of case definitions, and segmented regression. What is new is how they are combined to answer a question none of them addresses on its own. First, the proposition in Section 3.2 states a checkable

condition, the absence of covariance between the clinical return and the agent's actions within the policy update interval, under which the weight placed on clinical outcomes has no effect on the expected learning direction of a coding agent; this turns the informal claim that AI will optimise for billing into a hypothesis that deployment logs can falsify. Second, D(C) is defined relative to an explicit reference standard and linked to sensitivity, specificity and clinical prevalence by an exact identity (Section 3.3), which shows that its level is fixed by the reference definition and that its change under a fixed definition is the quantity to monitor. Third, injected coding changes and placebo deployment dates in synthetic populations establish the detection limits of the markers before any real data are analysed (Sections 3.7 and 4.5).

3. Materials and Methods

The study combines a theoretical framework (Sections 3.1 to 3.5), a set of experimental markers with a minimal empirical programme (Section 3.6), and a pre-study on synthetic data (Section 3.7).

3.1. Three Levels of AI Adaptability

The reification feedback loop [5] links documentation requirements, coded minimal medical data sets, policy and AI training, and new documentation requirements. Under human conditions, this loop is constrained by annual code updates, multi-year policy cycles, and the cognitive limits of human coders.

We distinguish three levels of AI adaptability: (a) static models with periodic retraining, where human teams curate data and retrain at discrete intervals; (b) dynamic models with tool and memory adaptation, which update retrieval databases and prompt templates between inference episodes without altering core architecture; and (c) self-updating agents, which autonomously change their own policy, reasoning strategies and evaluation criteria in response to feedback [7]. Level (c) closes the loop: the system's outputs modify the data on which its next iteration learns.

3.2. Decision-Process Model

We formalise the escalation as a partially observable Markov decision process (POMDP) [24] in which a coding agent acts on a patient population. Every component of the model is specified. Its parameters are not estimated here; Table 2 names the data from which each could be estimated.

Table 2. Parameters of the decision-process model and data from which they could be estimated.

Parameter	Meaning	Data for Estimation
α	Acceptance rate of AI code suggestions	Deployment and audit logs of coding software
ρ	Retention of an existing code between periods	Longitudinal claims or electronic health records
λ_C	Rate at which present conditions are coded without AI	Pre-deployment linkage of claims with clinical evidence
κ_C	Change in treatment probability caused by a code	Treatment rates among coded patients before and after deployment
τ, σ^2	Delay and noise of the clinical feedback signal	Registry linkage; time to readmission or laboratory follow-up
w_a, w_c, γ	Reward weights and discount	Vendor configuration and documentation
Payer adaptation	Frequency and scope of recalibration and audit	Payer rule changes and recalibration cycles

State. At time t the state is $s_t = (x_t, z_t)$. The clinical state x_t records, for every patient i and condition C, whether the condition is present ($x_{i,C,t} \in \{0, 1\}$) together with its severity. The administrative state z_t records which diagnosis codes are active in the patient's record ($z_{i,C,t} \in \{0, 1\}$).

Actions. At each step the agent chooses, for every patient–condition pair, whether to propose a code, propose its removal, or leave the record unchanged, $a_{i,C,t} \in \{+1, -1, 0\}$. The policy $\pi_\theta(\mathrm{a} \mid \mathrm{o})$ is parameterised by θ.

Transitions. The transition kernel factorises into a clinical and an administrative part, $\mathrm{P}(s_{t+1} \mid s_t, a_t) = P_x(x_{t+1} \mid x_t, z_t) \cdot P_z(z_{t+1} \mid z_t, x_{t+1}, a_t)$. Clinical dynamics depend on actions only through the record, because a coded diagnosis changes the probability of treatment along the care pathway; κ_C denotes the increase in treatment probability that a code produces. In the administrative kernel, a proposed code is accepted by the human reviewer with probability α, an existing code is retained from one period to the next with probability ρ, and codes also arise from ordinary clinical documentation at rate λ_C for patients in whom the condition is present.

Observation. The agent observes $o_t = (z_t, y_{t-\tau})$: the full administrative record, and a clinical signal $\mathrm{y} = \mathrm{g(x)} + \varepsilon$ that arrives with delay τ and noise $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, such as readmissions, laboratory follow-up or patient-reported outcomes. The observation function $\mathrm{O}(o_t \mid s_t)$ is therefore exact for the administrative component and delayed and noisy for the clinical component. This asymmetry is the core of the model.

Reward. The reward is $r_t = w_a\, R_a(z_t) + w_c\, R_c(x_t)$, with $w_a, w_c \geq 0$, where R_a is the administrative return of the coded record (reimbursement, quality-metric attainment) and R_c the clinical value of the care delivered. The agent receives $R_a(z_t)$ without delay but only an estimate of the clinical component, $\hat{R}_{c,t} = R_c(x_{t-\tau}) + \eta_t$. Returns are discounted with factor γ.

Learning. A level (c) system updates its own policy from experience. We take a policy-gradient update as the representative case [25], $\theta_{t+1} = \theta_t + \beta\, \hat{\nabla} J$, with $\hat{\nabla} J = \nabla \log \pi_\theta(\mathrm{a} \mid \mathrm{o})\, (w_a\, G_a + w_c\, \hat{G}_c)$, where G_a and $\hat{G}_c$ are the administrative and the estimated clinical return that follow the action.

Proposition. Suppose that, over the interval between policy updates, the estimated clinical return $\hat{G}_c$ is uncorrelated with the score $\nabla \log \pi_\theta(\mathrm{a} \mid \mathrm{o})$. Delay and noise do not by themselves imply this assumption. Delay implies it only when the clinical consequences of an action arrive after the update in which the action is evaluated and are not credited back to it. Noise that is independent of the action leaves the covariance unchanged in expectation and acts on the variance of the estimated update instead, which is the case treated in the last sentence of this paragraph. Then $\mathrm{E}[\hat{\nabla} J] = w_a\, \mathrm{E}[\nabla \log \pi_\theta\, G_a]$: the expected update follows the administrative objective alone, for any w_c, provided $w_a > 0$. *Derivation.* $\mathrm{E}[\nabla \log \pi_\theta\, \hat{G}_c] = \mathrm{Cov}(\nabla \log \pi_\theta, \hat{G}_c) + \mathrm{E}[\nabla \log \pi_\theta]\, \mathrm{E}[\hat{G}_c]$. The first term is zero by assumption, and the second is zero because the score has expectation zero under the policy. If the covariance is positive but small, the update combines both objectives, and the administrative direction dominates while $w_a\, \|\nabla J_a\|$ exceeds w_c times the clinical signal-to-noise ratio times $\|\nabla J_c\|$.

Hypothesis. The proposition is conditional: it shows what follows if the clinical return does not covary with the agent's actions at the timescale of learning. Whether that condition holds in deployed coding systems is an empirical question that the proposition does not answer. We therefore state the general claim as a hypothesis. H1: in deployed self-updating coding systems, the clinical signal available at the timescale of policy updates is too delayed or too noisy to counterbalance the administrative signal, so that coding drifts toward administrative optimisation. H1 is falsified if $\Delta \mathrm{D(C)}/\Delta \mathrm{t}$ after deployment does not differ from matched sites without such systems (Section 3.6).

Stationary or adaptive reward. The proposition treats R_a as fixed. In practice the payer responds: diagnosis-related group weights are recalibrated, risk-adjustment schemes apply coding-intensity adjustments, and audits target coding patterns that have become conspicuous [22,23]. The administrative reward is then the outcome of a repeated game between coding agent and payer rather than a stationary target. In addition, the agent's own

codes alter the data from which future rewards and training labels are computed, which is the setting of performative prediction [26]. Two consequences follow. Payer adaptation can remove the gain from a specific, detected coding pattern, but it does not remove the observability asymmetry, so the gradient is redirected toward patterns that are not yet audited, as Goodhart's law describes for optimised proxies [27], rather than eliminated. And the divergence is then expected to move between conditions over time rather than to grow in a single one, so that monitoring must cover a set of sentinel conditions.

POMDPs are established in healthcare for treatment optimisation, where they model hidden disease states with therapeutic actions. The inversion here is distinctive: the hidden state is clinical reality, the actions are documentary rather than therapeutic, and the agent's actions modify the observation space on which it subsequently learns.

3.3. The Artificial Epidemiology Divergence

The term artificial epidemiology and the notation $D(C) = P_a(C) - P_c(C)$ were introduced in [5] to designate the biased population health picture that can emerge when AI systems learn from data shaped by clinical–administrative negotiation. Here we define D(C) precisely, relate it to established measures, and specify how it is measured.

Definition 1. *For a clinical condition C and a reference standard R, the artificial epidemiology divergence is $D(C \mid R) = P_a(C) - P_c(C \mid R)$, where $P_a(C)$ is the administrative prevalence (patients with an active code from a specified code set) and $P_c(C \mid R)$ is the clinical prevalence (patients who meet R). R fixes the clinical criteria, the treatment criteria and the time window.*

$D(C) > 0$ indicates administrative over-representation and $D(C) < 0$ administrative under-representation relative to R. Because P_c is defined by R, every reported value of D(C) is accompanied by its reference definition.

Relation to code validation. If Se and Sp denote the sensitivity and specificity of the code against R, then $P_a = \mathrm{Se} \cdot P_c + (1 - \mathrm{Sp})(1 - P_c)$, and hence $D(C \mid R) = (1 - \mathrm{Sp})(1 - P_c) - (1 - \mathrm{Se})\, P_c$. In counts, $D(C \mid R)$ equals the number of patients who are coded but do not meet R, minus the number who meet R but are not coded, divided by the population size. D(C) is therefore a net summary of the same two-by-two table that validation studies of administrative case definitions report against chart review [20,21,28]. It differs from those studies in three respects. First, it is a population-level net quantity in which errors in opposite directions cancel; it is therefore always reported together with its two components. Second, it is designed for repeated measurement under a fixed reference. An offset that arises from the choice of R, for example, from comparing a diagnosis code with a single laboratory threshold, cancels in $\Delta D(C \mid R)/\Delta t$ only while it is constant over time, whereas a change in coding behaviour does not cancel. A fixed reference definition does not by itself keep the offset constant: by the identity above, it depends on P_c, Se and Sp, which change when the age structure or morbidity of the population changes or when testing patterns change, since a laboratory-based reference counts only those who are tested. Changes in $D(C \mid R)$ are therefore interpreted together with the testing rate and the population structure, standardised where these change, and compared with matched sites. Third, it differs from coding-intensity measures, which compare coding over time or between providers against expected morbidity [6,22,23]: D(C) requires linked clinical evidence and in exchange separates more coding from more disease.

Ascertainment. P_c can only count patients who were tested, and testing may itself follow a diagnosis code. The share of the population with a qualifying measurement is therefore reported with every estimate.

Illustrative example. Hang et al. found that administrative coding of suicide attempts in an Australian emergency minimum dataset achieved 25% sensitivity against medical

record review [29]. If the clinical prevalence of documented suicide attempts were 2.5% and false-positive coding negligible, administrative prevalence would be about 0.625%, yielding D(C) = −1.875 percentage points. The example uses published sensitivity data rather than original analysis and shows why D(C) is used as a monitoring indicator, tracking $\Delta D(C)/\Delta t$ around deployment, rather than as a prevalence estimate.

Without self-updating AI, D(C) changes at human speed, driven by annual code updates, policy revisions, and shifting institutional incentives. With self-updating AI, D(C) may shift continuously, asymmetrically and without external visibility, as the system updates its policy between encounters.

3.4. Strange Loops and Requisite Variety

Ashby's law of requisite variety provides the information-theoretic constraint: effective regulation requires that the regulator's variety matches the variety of the system being regulated [30]. Standardised coding reduces variety, compressing clinical experience into a finite set of categorical codes. A self-updating system operating within this reduced variety space optimises for the categories that exist, concentrating its outputs on well-coded, well-reimbursed conditions. Conditions lacking adequate categorical representation can become progressively invisible; this is the mechanism through which $D(C) < 0$ conditions are expected to worsen.

Hofstadter's strange loop describes the self-referential structure [31]: AI trained on coded data produces documentation that becomes coded data that trains the next iteration. Luhmann's concept of autopoiesis describes systems that reproduce the conditions of their own existence through their own operations [32]. Zönnchen et al. examined whether large language models can be understood as autopoietic in Luhmann's sense and concluded that the frame is productive but analogical rather than literal [33]; we adopt their caution and do not resolve whether the healthcare case constitutes autopoiesis in Maturana and Varela's biological sense [34]. The empirical predictions in Section 3.6 rest solely on the observability asymmetry (Section 3.2) and the recursive training mechanism [18].

3.5. Threats to Validity

T1. Confounders. Secular coding trends, regulatory changes such as annual code revisions and diagnosis-related group recalibration, and demographic shifts could produce D(C) patterns indistinguishable from AI-driven artificial epidemiology. Any empirical test requires matched non-AI control sites and interrupted time-series designs.

T2. Marker non-specificity. Code distribution convergence and comorbidity inflation may reflect guideline revisions, new screening programmes, or changing disease prevalence. The falsification conditions in Section 3.6 address this, and Section 4.5 reports the size of change detectable against variation without intervention in synthetic data.

T3. Reference availability. For conditions without an independent clinical reference (many primary care presentations, functional somatic syndromes, subclinical mental health conditions), D(C) cannot be computed. These are likely among the conditions most vulnerable to artificial epidemiology.

T4. Reference dependence. D(C) depends on the reference definition R (Section 4.3). Comparisons are valid only under a fixed R, and the baseline value carries no information about coding quality on its own.

T5. Non-parameterisation. The decision-process model is specified but not parameterised. It yields a conditional statement and a hypothesis, not quantitative predictions.

3.6. Experimental Markers and Minimal Empirical Programme

We propose five markers that can be tested with existing data infrastructure (claims databases, clinical registries, AI deployment logs). For each marker we state the measurement, the expected observation under H1, and a falsification condition.

Marker 1: Code distribution convergence. Measured as the Jensen–Shannon divergence (base 2, bounded between 0 and 1) between the distribution of a fixed set of codes in period t and in a pre-deployment reference period. The Kullback–Leibler divergence is asymmetric and undefined when a code that appears after deployment is absent from the reference period; we therefore use its symmetric, bounded form, with a constant of 0.5 added to every cell. Under H1 the distribution should shift toward high-reimbursement codes, as the model-collapse findings predict [18]. *Control and falsification:* the same shift in matched non-AI sites (difference in differences), or at placebo deployment dates.

Marker 2: Comorbidity inflation. The ratio of observed to expected co-occurrence for billing-relevant code pairs. Under H1 the ratio should rise after AI deployment without corresponding changes in clinical evidence. *Falsification:* clinical evidence confirms a genuine increase in comorbidity.

Marker 3: Documentation-action gap. The change in treatment rates among coded patients, measured against matched sites or the site's own pre-deployment period, not the absolute share of untreated coded patients. Absence of treatment is not evidence of miscoding: lifestyle management, contraindications, controlled disease and patient preference all explain untreated diagnoses, and treatment without a code also occurs (Section 4.4). Under H1, coded diagnoses should increase while treatment rates among the coded fall, as observed by Castro et al. [1] and in the diagnosis–treatment discordance reported in [2]. *Falsification:* documentation and treatment change proportionally.

Marker 4: Equity divergence. The difference in D(C | R) between population subgroups. Under H1, D(C) should widen asymmetrically, with larger positive D(C) in well-resourced populations and larger negative D(C) in underserved populations. *Falsification:* D(C) shifts uniformly across subgroups.

Marker 5: Temporal acceleration. The change in the slope of D(C | R) at deployment, estimated by segmented regression. Under H1 the slope should change after deployment. *Falsification:* the change in slope does not exceed that at placebo dates or at matched sites.

Three study designs could operationalise these markers. (1) A before–after analysis at a site deploying AI-assisted coding compares code distributions, comorbidity co-occurrence, treatment rates and D(C) in the periods before and after go-live, using an interrupted time series. (2) Where several sites adopt AI coding at different times, an event-study design uses staggered adoption as a natural experiment, with sentinel conditions that have laboratory or registry evidence (diabetes, hypertension, cancer, acute myocardial infarction) as anchors. (3) A cross-site comparison of AI-adopting and matched non-AI sites tests whether the markers change more in AI sites. Designs (1) to (3) can be conducted with pseudonymised linked records or aggregate statistics within the secondary-use provisions of the European Health Data Space [35].

3.7. Pre-Study on Synthetic Populations

The pre-study had two aims: to measure D(C) and the markers in data that contain no AI coding process, which shows the baseline against which any AI effect would have to be detected, and to determine how large an injected coding change must be before it is detectable. It does not test H1.

Rationale for synthetic data. Synthetic electronic health records are an accepted means of developing and testing analytic methods before access to clinical data is granted, and the validity of Synthea for this purpose has been examined against clinical quality measures [36].

Real-world claims linked with clinical evidence, which are needed to compute D(C) directly, require institutional data access agreements that were not available for this study. Synthetic data also offer what real data cannot: a known generating process without billing incentives or AI coding. Every divergence they contain is therefore a property of the measurement construct, which makes them the appropriate negative control for the markers, and the synthetic baseline can later be compared with real data to show where the generator departs from clinical practice.

Populations. We used Synthea [37,38], master branch, commit aa0772fb of 5 March 2026, with all modules enabled, a reference date of 22 March 2026 and ten years of exported history. The primary population (Massachusetts, seed 1774160093279) comprises 11,475 patients. Eight further populations of 3000 living patients each (about 3400 including deceased) were generated with seeds 101 to 104 for Massachusetts and 201 to 203 for Texas, Ohio and Washington. One Texas population was generated twice with the same seed on two computers (Java 17 and Java 21) to quantify variation that arises from the generator alone: the two runs shared 3334 of 3342 patient identifiers but were not identical. The nine datasets therefore comprise eight distinct populations and one repeated generation with the same seed; they are not nine independent replications, and counts and ranges across the nine datasets are descriptive. All data are synthetic.

Cohort. For each population, the analysis cohort comprised patients aged 18 years or older who were alive at the reference date and had at least one encounter in the three preceding years.

Code sets. Administrative prevalence P_a(C) counted patients with a code from an explicit SNOMED CT code set that was active in the window (onset before the reference date and no resolution before the start of the window). The code sets are listed in Table S1: diabetes comprised type 1 and type 2 diabetes mellitus and their complications, and excluded prediabetes, which was analysed as a separate condition.

Reference definitions. The primary reference R combined laboratory evidence and treatment (Table 3). Diabetes: last HbA1c in the window $\geq$ 6.5%, any HbA1c $\geq$ 6.5% before the reference date, or an active glucose-lowering drug other than metformin (metformin was excluded because Synthea prescribes it for prediabetes). Prediabetes: last HbA1c 5.7% to 6.4% in the window, without meeting the diabetes definition. Hypertension: last systolic blood pressure $\geq$ 140 mmHg or diastolic $\geq$ 90 mmHg, or an active antihypertensive drug from the angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, thiazide or calcium channel blocker classes. Hyperlipidaemia: last total cholesterol $\geq$ 240 mg/dL or LDL cholesterol $\geq$ 160 mg/dL, or an active statin. Obesity: last body mass index $\geq$ 30 kg/m^2. Sensitivity analyses varied the definition (laboratory value only; two values above the threshold), the threshold (Table 3), the window (one and five years), and the administrative definition (codes ever recorded).

Table 3. Reference definitions for the clinical prevalence P_c(C | R).

Condition	Primary Reference R	Threshold Sensitivity
Diabetes	Last HbA1c $\geq$ 6.5% in window, or any HbA1c $\geq$ 6.5% before, or glucose-lowering drug other than metformin	HbA1c 6.0%, 7.0%
Prediabetes	Last HbA1c 5.7% to 6.4% in window, not meeting the diabetes reference	none
Hypertension	Last systolic $\geq$ 140 or diastolic $\geq$ 90 mmHg, or antihypertensive drug	130/80, 150/95 mmHg
Hyperlipidaemia	Last total cholesterol $\geq$ 240 or LDL $\geq$ 160 mg/dL, or statin	200/130, 280/190 mg/dL
Obesity	Last body mass index $\geq$ 30 kg/m^2	27, 35 kg/m^2

Analyses. For each condition, we report P_a, P_c, D(C) with a 95% bootstrap interval (1000 patient-level resamples), the coded-only and clinical-only counts, the sensitivity and positive predictive value of the code against R, and the share of the cohort with a qualifying

measurement. Coded-only patients were further classified as having met the threshold only before the window, having no measurement in the window, or having been measured below the threshold. Marker 2 was computed as observed over expected co-occurrence for all pairs of the five conditions. Marker 3 was computed as the share of patients with a diabetes code without any glucose-lowering drug, stratified by last HbA1c, and correspondingly for hypertension and hyperlipidaemia. Marker 4 was computed as D(C) by race and ethnicity and by income tertile. Markers 1 and 5 were computed on yearly data for 2017 to 2025, the full calendar years within the export window: D(C) per year for the cohort with an encounter in that year, with a three-year laboratory window, and the Jensen–Shannon divergence of the sentinel code distribution against the first year.

Injected coding scenarios. To test detectability, two coding changes were injected from a deployment year of 2021. In the upcoding scenario, each year a fraction f of cohort patients with an HbA1c of 5.7% to 6.4% and no diabetes code received a diabetes code, which was retained in later years. In the undercoding scenario, each year a fraction f of patients coded with hypertension lost the code. Fractions were 5%, 10% and 20% per year. The detection statistic was the difference between the post- and pre-deployment slopes of D(C), and the null distribution was formed by the same statistic at placebo deployment years (2019 to 2022) in the data without injection, across all populations. A scenario counted as detected in a population when its slope change exceeded the largest absolute placebo value.

Software and reproducibility. Analyses used Python 3.14.2 with pandas 3.0.0 and NumPy 2.4.2 for six populations and Python 3.12.3 with pandas 3.1.0 and NumPy 2.5.4 for the three datasets analysed on the second computer (Ohio, Washington and the Texas generation with Java 21); the random seed for bootstrap and scenarios was 20261029, with one generator per population. Code, code sets and aggregate results are provided in the Supplementary Materials (Data Availability Statement).

Use of generative AI. Generative AI tools (Claude, Anthropic) were used for literature search support, formatting assistance and generation of the analysis code. All scientific content, theoretical arguments, the decision-process formalisation, the design of the D(C) metric and clinical interpretation were produced by the authors, who take full responsibility for the work.

4. Results

4.1. Populations

The primary analysis cohort comprised 7836 adults; the eight further cohorts comprised 2222 to 2390 adults each.

4.2. Divergence Under the Primary Reference

Table 4 reports D(C) under the primary reference. Diabetes showed a positive divergence of +0.041, hypertension and hyperlipidaemia negative divergences of −0.125 and −0.191, and obesity the largest positive divergence, +0.314. The sign of D(C) was the same in all nine datasets for every condition, and the range across populations was narrow relative to the differences between conditions.

The components show where the divergences come from. Of the 319 patients coded with diabetes who did not meet the reference, 173 had no HbA1c measurement in the window and 146 had been measured below 6.5% without any value at or above it; the diabetes module of the generator assigns the diagnosis through criteria that a single HbA1c threshold does not capture. Only 48% of the cohort had an HbA1c measurement in the window so that P_c for diabetes and prediabetes is bounded by ascertainment. Of the 2515 patients coded with obesity who did not meet the reference, 1380 had a body mass index of 30 kg/m^2 or more before the window and 1135 had a last value below 30 kg/m^2 in the

window: the code was retained after the measured value fell, which corresponds to the retention parameter ρ of the model. Of the 979 clinical-only patients for hypertension, 826 met the reference through a single last blood pressure reading alone; of the 1495 clinical-only patients for hyperlipidaemia, 848 received a statin without a hyperlipidaemia code, consistent with preventive prescribing.

Table 4. Artificial epidemiology divergence under the primary reference definition. Primary population (Massachusetts, n = 7836), with the range of D(C) across all nine populations.

Condition	P_a	P_c	D(C) (95% CI)	Coded Only	Clinical Only	Se of Code	PPV of Code	Tested	D(C), Range of 9 Populations
Diabetes	0.153	0.112	+0.041 (0.036 to 0.045)	319	0	1.00	0.73	48%	+0.035 to +0.049
Prediabetes	0.450	0.354	+0.096 (0.089 to 0.103)	782	29	0.99	0.78	48%	+0.087 to +0.100
Hypertension	0.253	0.378	−0.125 (−0.132 to −0.118)	0	979	0.67	1.00	100%	−0.132 to −0.105
Hyperlipidaemia	0.111	0.301	−0.191 (−0.199 to −0.182)	0	1,495	0.37	1.00	73%	−0.213 to −0.174
Obesity	0.569	0.255	+0.314 (0.303 to 0.324)	2,515	54	0.97	0.44	100%	+0.294 to +0.334

Se, sensitivity; PPV, positive predictive value; Tested, share of the cohort with a qualifying measurement in the window.

4.3. Dependence on the Reference Definition

Figure 1 shows D(C) under each alternative reference definition in all populations. For diabetes, D(C) ranged from −0.321 to +0.129 across definitions, and became strongly negative at an HbA1c threshold of 6.0%, which counts most patients with prediabetes as clinical cases. For hypertension, D(C) was negative under the primary reference but positive when only laboratory values were used (+0.042 to +0.075), because the primary reference counts treated patients and the laboratory-only reference does not; across thresholds it ranged from −0.427 to −0.035. Hyperlipidaemia changed sign in the same way. For obesity, D(C) ranged from −0.257 to +0.553 between thresholds of 27 and 35 kg/m^2. The time window had little influence. The baseline value of D(C) is therefore a property of the pair of the code set and reference definition, not of the population alone.

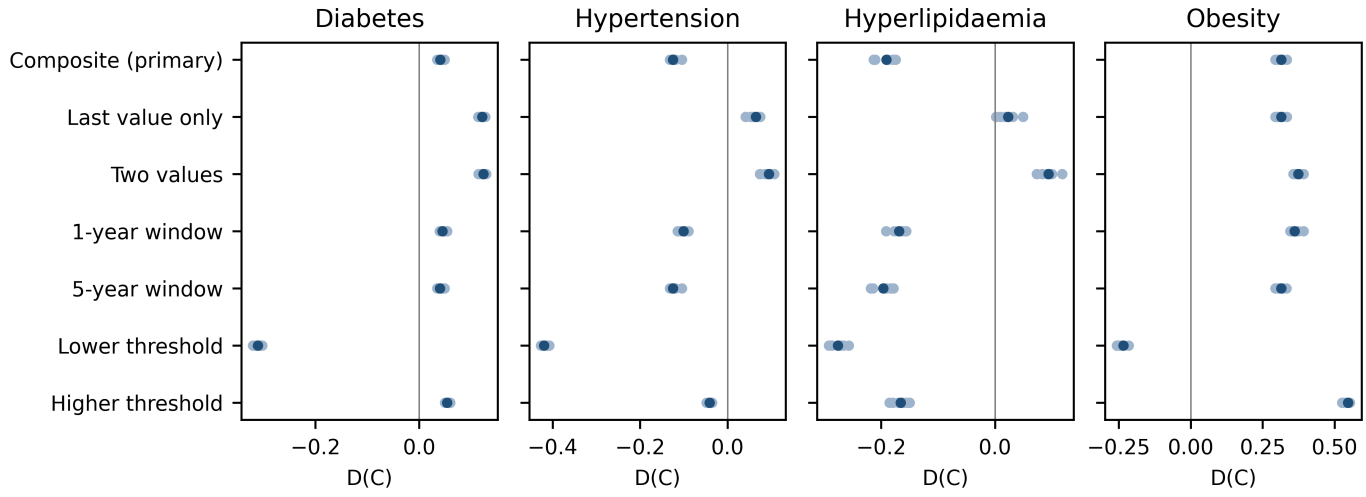


Figure 1. D(C) under alternative reference definitions for four conditions. Each point is one population; the primary population is shown in dark blue. The composite reference is the primary definition; "last value only" and "two values" use laboratory or vital-sign evidence without treatment; lower and higher thresholds vary the cut-off of the composite reference as listed in Table 3.

4.4. Markers 2 to 4 in the Absence of AI Coding

Comorbidity (Marker 2). All ten pairs of conditions co-occurred more often than expected under independence, with ratios from 1.24 to 3.30; the highest was diabetes with hypertension (2.92 to 3.30). In data without coding incentives this reflects the generator's disease clustering and establishes the baseline that a change after deployment would have to exceed.

Documentation-action gap (Marker 3). In the primary population, 388 of 1198 patients coded with diabetes (32.4%; 32.1% to 36.6% across populations) had no active glucose-lowering drug. Of these, 201 had a last HbA1c of 6.5% or more, 14 a last value below 6.5%, and 173 no measurement in the window. Every patient coded with hypertension received an antihypertensive drug, and all but one coded with hyperlipidaemia received a statin. In the generator, the untreated coded diabetes cases arise from its care pathways, not from miscoding, which illustrates why Marker 3 is defined as a change rather than as an absolute share.

Equity divergence (Marker 4). Differences in D(C) between subgroups were moderate relative to the differences between conditions. Across populations, the span between the highest and lowest subgroup D(C) (groups of at least 200 patients) was at most 0.080 by race and ethnicity and at most 0.087 by income tertile. Synthea models some demographic differences in disease incidence but no differences in coding behaviour, so these values represent the subgroup heterogeneity that exists without any coding bias.

4.5. Markers 1 and 5: Detection of Injected Coding Change

Without injection, the yearly D(C) for diabetes changed little between 2017 and 2025, and the absolute change in slope at placebo deployment years did not exceed 0.0026 per year for diabetes and 0.0116 per year for hypertension across all populations (Figure 2). Injected upcoding of prediabetes to diabetes changed the slope of D(C) by 0.013 to 0.017 per year at 5% per year, 0.022 to 0.026 at 10% and 0.035 to 0.041 at 20% and was detected in 9 of 9 populations at 5% per year. Injected loss of hypertension codes changed the slope by −0.010 to −0.004 at 5% per year, which did not exceed the largest placebo value (0 of 9 populations detected), although the cumulative departure from the populations without injection was visible by 2025 (Figure 2), and by −0.021 to −0.014 at 10% per year, detected in 9 of 9 populations. The Jensen–Shannon divergence of the code distribution in 2025 against 2017 was 0.0007 to 0.0014 bits without injection and rose to 0.0204 to 0.0251 bits under 20% upcoding and to 0.0016 to 0.0028 bits under 5% undercoding.

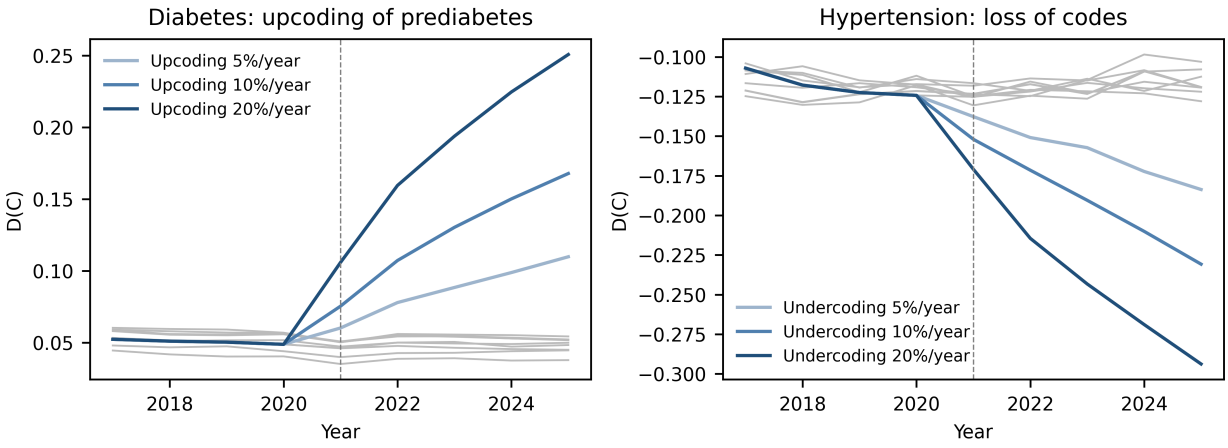


Figure 2. Yearly D(C) from 2017 to 2025. Grey lines, all populations without injected coding change; coloured lines, primary population with injected upcoding of prediabetes to diabetes (**left**) or loss of hypertension codes (**right**) at 5%, 10% and 20% per year from 2021 (dashed line).

5. Discussion

5.1. What the Synthetic Pre-Study Shows and What It Cannot Show

The pre-study separates two kinds of divergence that the framework must keep apart. The first is construct divergence: the gap between a diagnosis code and a laboratory- or treatment-based reference that exists without any AI or billing incentive. In the synthetic populations it arises mechanically from the difference between single-measurement thresholds and the multifactorial logic by which the generator assigns diagnoses, from the retention of codes after measured values change, from ascertainment limited to those tested, and from treatment given for indications other than the coded condition. Its size and even its sign depend on the reference definition. The second is change in coding behaviour, which the framework attributes to administrative optimisation and which, in real data, would add to the construct divergence. The synthetic data contain only the first kind. The divergences reported here therefore say nothing about AI, about billing, or about the financially motivated negotiation of codes that is the substance of the theory. They show what any comparison of codes with clinical evidence will find before AI enters the picture, and that the second kind becomes detectable against this baseline once it exceeds a certain size.

This separation answers the principal methodological objection to D(C) as a level: a positive or negative D(C) at one point in time cannot be read as evidence of upcoding or undercoding. The quantity that carries the information is the change in D(C) under a fixed reference, which removes the construct offset as long as population characteristics and testing patterns remain stable (Section 3.3). In the synthetic data, an upcoding of 5% of eligible patients per year was detectable within a few years, while a loss of 5% of codes per year was not; the detection limit is asymmetric because the two scenarios act on populations of different size and against different background variation. Real data will be noisier than synthetic data, so these limits are lower bounds on the effect sizes that real studies would need.

The payer analysis described in the Introduction [2] provides a large-scale real-world observation of the documentation-action pattern in claims, but, as its critics note [3], claims alone cannot separate legitimate from artefactual diagnoses. Computing the change in D(C) on linked administrative and clinical records, at sites that deploy AI coding and at matched sites that do not, is the step that would resolve that dispute.

5.2. When the Feedback Loop Does Not Apply

Three conditions may attenuate or prevent artificial epidemiology. Clinical registries with independent data collection, such as cancer registries and disease-specific cohorts, maintain P_c(C) independently of billing processes, making D(C) monitorable. Single-payer systems with salaried physicians and clinical coding audits attenuate the financial incentive, although documentation burden and diagnostic pressure remain. AI systems built on dual-track architectures such as openEHR [39] maintain separate clinical and administrative layers. Fairness constraints and bias detection [12,13] are sufficient while the loop is open (static AI with human coding review); they become insufficient when the loop closes, because the bias is then a dynamic property of the system rather than a fixed property of the training data. Architectures that permanently eliminate uncertain hypotheses are structurally aligned with administrative optimisation, whereas architectures that preserve them through strategic dormancy, as in the EMoT reasoning architecture [40], maintain the variety needed for clinical-first governance.

5.3. Conditions Under Which the Framework Would Be Wrong

The framework rests on falsifiable premises. If clinical feedback is timely and precise enough at the timescale of policy updates, the condition of the proposition fails and H1 does not follow. If self-updating clinical AI systems remain constrained to human-reviewed

rule sets, so that their outputs never feed back into their own training data, the loop does not close. If production systems are retrained only on fresh, externally curated data, the recursive degradation mechanism [18] does not apply. And if, in linked real-world data, the change in D(C) after AI deployment does not exceed that at matched sites, H1 is falsified. Under any of these conditions, the existing toolkit of code validation, coding audit and bias detection would remain adequate.

5.4. Implications for the European Health Data Space

The European Health Data Space regulation establishes cross-border secondary use of health data, explicitly including AI training; its provisions on secondary use apply from 26 March 2029, with some data categories following in 2031 [35]. It addresses data quality in terms of completeness, consistency and timeliness but does not distinguish between clinical and administrative data quality. D(C) may vary substantially by member state: a statutory health insurance system with strong billing incentives may produce different D(C) profiles from a single-payer system. Health Data Access Bodies could require that datasets made available for secondary use report D(C | R) for sentinel conditions with a stated reference definition so that users can see how coded and clinical prevalence relate and whether that relation changes. In systems where reporting targets rather than reimbursement shape coding, the relevant asymmetry may involve performance indicators rather than financial incentives, but the structural logic may still hold.

5.5. Implications for General Practice

A solo practitioner in a rural setting codes under different constraints than a physician in a managed centre with dedicated coding staff. A substantial proportion of office-based physicians in Germany still practise in solo or small-group settings [41], while larger centres increasingly deploy AI-assisted documentation with standardised coding pipelines. If AI decision support is sensitive to coding framing, the quality of AI-assisted care may vary with the administrative infrastructure around the physician rather than with clinical competence. The framework offers a way to measure whether this structural difference translates into differential coding drift.

From the practitioner's perspective, the most actionable point is not the existence of coding-system bias, which most experienced general practitioners would consider unsurprising, but its invisibility to the clinician using the tool. A physician receiving an AI-generated suggestion cannot know whether it was influenced by the coding system in which the patient's history was recorded. Governance responses therefore need an architectural component: certification of clinical AI could include coding-robustness testing, analogous to the bias and data-governance requirements for high-risk AI systems under the EU AI Act [42], and monitoring of the change in D(C) could serve as one component of such testing.

5.6. Limitations

First, the decision-process model is specified but not parameterised, and its central claim is a conditional proposition and a hypothesis, not a demonstrated property of deployed systems. Second, the pre-study uses synthetic data without any AI coding process; it establishes the measurement approach, its baseline and its detection limits, and cannot show that AI produces prevalence divergence. Third, the generator assigns diagnoses through its own clinical modules, so that construct divergences in synthetic data need not resemble those in real records, and the injected scenarios are deliberately simple. Fourth, the synthetic populations are not representative of any real population. They come from one generator, whose demographic inputs and disease modules are based on US data and clinical guidelines, and from four US states; they do not reflect the age structure, care pathways, coding systems or reimbursement rules of European health systems, nor the documentation habits of individual

clinicians and practices. The values of D(C) reported here are therefore not estimates for any real population but a negative control for the measurement approach. In addition, seeds reproduce populations closely but not identically across computing environments. Fifth, the clinical references use laboratory values, vital signs and drug classes and therefore inherit their limits: treated patients with normal values, preventive prescribing, and conditions that are measured only in those tested. Sixth, conditions without independent clinical evidence cannot be monitored with D(C) (T3). Seventh, the financial incentive component applies most to systems with strong billing incentives; in low- and middle-income countries the loop may operate through public health reporting or donor-driven data requirements. Eighth, self-updating AI systems in healthcare are research prototypes [8,9], and the argument about their deployment is prospective. Finally, the payer analysis cited [2] is a claims-based report by an interested party and has been contested [3]; it is used here as an observation of the pattern, not as validation of the framework.

5.7. Future Work

Four lines of work follow from this study. First, the change in D(C | R) should be computed on linked administrative and clinical records around the deployment of AI-assisted coding, using the before–after, staggered-adoption and matched-site designs in Section 3.6, with sentinel conditions for which laboratory or registry evidence exists; under the European Health Data Space, Health Data Access Bodies could provide such linkage for secondary use [35]. Second, the decision-process model should be parameterised: the delay and noise of clinical feedback, the policy update interval of deployed coding systems and the covariance between clinical outcomes and coding actions within that interval can be estimated from deployment logs, which allows for the condition of the proposition to be tested directly. Third, the synthetic generator could be extended with explicit coding-behaviour modules, including billing incentives and a learning coding agent, so that the full feedback loop is simulated and the markers are tested against a known amplification process. Fourth, the analysis should be extended beyond metabolic conditions to conditions with registry evidence, such as cancer and acute myocardial infarction, and to European populations, where incentives differ between statutory insurance and single-payer systems.

6. Conclusions

The reification feedback loop describes how administrative categories enter clinical data through ordinary documentation practice [5]. Self-updating AI could turn this slow institutional drift into a faster, self-reinforcing process. The decision-process model shows that, if the clinical return does not covary with the agent's actions at the timescale of learning, policy-gradient learning follows the administrative reward regardless of the weight placed on clinical outcomes; delayed or noisy clinical feedback can bring this condition about but does not guarantee it, and whether that condition holds in practice is a hypothesis to be tested, not a result. D(C) provides a measurable construct for the population-level consequences, provided it is reported with its reference definition and interpreted as a change rather than a level. The synthetic pre-study shows that construct divergence is present without any AI, that it depends on the reference definition, and that coding changes of a few per cent per year become detectable against this baseline. The study establishes a measurement approach; it does not show that self-updating AI causes prevalence distortion. The practical next step is to compute the change in D(C) for sentinel conditions where AI-assisted coding is being deployed, with a fixed reference definition and matched sites without AI coding.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/medsci14070777/s1. Table S1: SNOMED CT code sets and drug classes. Table S2: D(C) under all reference definitions for all populations. Table S3: Subgroup D(C) and documentation-action results. File S1: Analysis code, run scripts and aggregate results with a README describing data generation and the order of execution.

Author Contributions: Conceptualization, F.O.S., H.T. and T.F.; Methodology, F.O.S.; Software, F.O.S.; Validation, H.T.; Formal Analysis, F.O.S.; Visualization, F.O.S.; Writing—Original Draft Preparation, F.O.S.; Writing—Review & Editing, F.O.S., H.T. and T.F.; Supervision, T.F.; Project Administration, F.O.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study used synthetic data only and did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article and its Supplementary Materials; the analysis code, code sets, run scripts and aggregate results are provided in File S1. The synthetic records were generated with Synthea (master branch, commit aa0772fb of 5 March 2026) using the parameters stated in Section 3.7; they contain no real patient information and are available from the corresponding author on request.

Acknowledgments: The authors thank colleagues across the WONCA Europe and EGPRN networks for exchange on how clinical AI deployments interact with administrative coding in European general practice. A preprint of an earlier version of this work was posted on Cambridge Open Engage [4]. During the preparation of this study, the authors used Claude (Anthropic; Claude Sonnet 4.6, Claude Opus 4.6 and Claude Opus 5.5) for the purposes of literature search support, formatting assistance and generation of the analysis code. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: F.O.S. holds equity and operational interests in primary care AI products that are not the subject of this paper and are not cited in it. T.F. is President of WONCA Europe; this role involves no financial interests in the subject of the paper. The authors declare no other conflicts of interest.

References

1. Castro, V.M.; McCoy, T.H.; Verhaak, P.; Ramachandiran, A.; Perlis, R.H. Psychiatric documentation and management in primary care with artificial intelligence scribe use. *JAMA Psychiatry* **2026**, *83*, 281–286. https://doi.org/10.1001/jamapsychiatry.2025.4303.
2. Birkmeyer, C.; Wennberg, D.; Kamons, K.; Chalker, L. Hospital Coding Intensity Analysis: Major Bowel Procedures. In *Evidence of Secondary Diagnosis Growth Driving DRG Coding Intensity*; Blue Cross Blue Shield Association: Chicago, IL, USA, 2026. Available online: https://www.bcbs.com/media/pdf/BCBSA-AI-Coding-Intensity-Whitepaper.pdf (accessed on 14 October 2026).
3. Smith, M.; Holloman, J. Better Information, Better Care: How Coding Supports Today's Complex Patients. *American Hospital Association*, 5 October 2026. Available online: https://www.aha.org/news/blog/2026-10-05-better-information-better-care-how-coding-supports-todays-complex-patients (accessed on 14 October 2026).
4. Stummer, F.O. Artificial epidemiology: How self-evolving clinical AI manufactures disease prevalence from administrative coding artifacts. *Camb. Open Engag.* **2026**, *preprint*. https://doi.org/10.33774/coe-2026-ssm1q.
5. Stummer, F.O. The symptom and the billing code: Documentary practices in minimal medical data sets as sites of clinical-administrative negotiation. *SSRN* **2026**, *preprint*. https://doi.org/10.2139/ssrn.6475038.
6. Crespin, D.; Dworsky, M.; Levin, J.; Ruder, T.; Whaley, C.M. Upcoding linked to up to two-thirds of growth in highest-intensity hospital discharges in 5 states, 2011–19. *Health Aff.* **2024**, *43*, 1619–1627. https://doi.org/10.1377/hlthaff.2024.00596.
7. Gao, H.-a.; Geng, J.; Hua, W.; Hu, M.; Juan, X.; Liu, H.; Liu, S.; Qiu, J.; Qi, X.; Wu, Y.; et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv* **2025**, arXiv:2507.21046.
8. Zhu, Y.; Qi, Y.; Wang, Z.; Gu, L.; Sui, D.; Hu, H.; Zhang, X.; He, Z.; He, J.; Ma, L.; et al. HealthFlow: A self-evolving AI agent with meta planning for autonomous healthcare research. *arXiv* **2025**, arXiv:2508.02621.
9. Jin, R.; Zhang, Z.; Wang, M.; Cong, L. STELLA: Towards a biomedical world model with self-evolving multimodal agents. *bioRxiv* **2025**, https://doi.org/10.1101/2025.07.01.662467.

10. Venkatesh, K.P.; Raza, M.M.; Kvedar, J.C. Automating the overburdened clinical coding system: Challenges and next steps. *npj Digit. Med.* **2023**, *6*, 16. https://doi.org/10.1038/s41746-023-00768-0.
11. Dong, H.; Falis, M.; Whiteley, W.; Alex, B.; Matterson, J.; Ji, S.; Chen, J.; Wu, H. Automated clinical coding: What, why, and where we are? *npj Digit. Med.* **2022**, *5*, 159. https://doi.org/10.1038/s41746-022-00705-7.
12. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. https://doi.org/10.1126/science.aax2342.
13. Chin, M.H.; Afsar-Manesh, N.; Bierman, A.S.; Chang, C.; Colón-Rodríguez, C.J.; Dullabh, P.; Duran, D.G.; Fair, M.; Hernandez-Boussard, T.; Hightower, M.; et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Netw. Open* **2023**, *6*, e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050.
14. Jiang, R.; Chiappa, S.; Lattimore, T.; György, A.; Kohli, P. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*; ACM: New York, NY, USA, 2019; pp. 383–390. https://doi.org/10.1145/3306618.3314288.
15. Ensign, D.; Friedler, S.A.; Neville, S.; Scheidegger, C.; Venkatasubramanian, S. Runaway feedback loops in predictive policing. *Proc. Mach. Learn. Res.* **2018**, *81*, 160–171.
16. Fries, J.A.; Steinberg, E.; Khattar, S.; Fleming, S.L.; Posada, J.; Callahan, A.; Shah, N.H. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **2021**, *12*, 2017. https://doi.org/10.1038/s41467-021-22328-4.
17. Bossen, C.; Pine, K.H. Batman and Robin in healthcare knowledge work: Human-AI collaboration by clinical documentation integrity specialists. *ACM Trans. Comput.-Hum. Interact.* **2023**, *30*, 1–29. https://doi.org/10.1145/3569892.
18. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755–759. https://doi.org/10.1038/s41586-024-07566-y.
19. Tempest-Walters, K. AI epidemiology: Achieving explainable AI through expert oversight patterns. *arXiv* **2025**, arXiv:2512.15783.
20. Quan, H.; Sundararajan, V.; Halfon, P.; Fong, A.; Burnand, B.; Luthi, J.-C.; Saunders, L.D.; Beck, C.A.; Feasby, T.E.; Ghali, W.A. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **2005**, *43*, 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83.
21. van Walraven, C.; Austin, P. Administrative database research has unique characteristics that can risk biased results. *J. Clin. Epidemiol.* **2012**, *65*, 126–131. https://doi.org/10.1016/j.jclinepi.2011.08.002.
22. Silverman, E.; Skinner, J. Medicare upcoding and hospital ownership. *J. Health Econ.* **2004**, *23*, 369–389. https://doi.org/10.1016/j.jhealeco.2003.09.007.
23. Geruso, M.; Layton, T. Upcoding: Evidence from Medicare on squishy risk adjustment. *J. Polit. Econ.* **2020**, *128*, 984–1026. https://doi.org/10.1086/704756.
24. Kaelbling, L.P.; Littman, M.L.; Cassandra, A.R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **1998**, *101*, 99–134. https://doi.org/10.1016/S0004-3702(98)00023-X.
25. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. https://doi.org/10.1007/BF00992696.
26. Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; Hardt, M. Performative prediction. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020), Virtual, 13–18 July 2020; Volume 119, pp. 7599–7609.
27. Manheim, D.; Garrabrant, S. Categorizing variants of Goodhart's law. *arXiv* **2018**, arXiv:1803.04585.
28. Benchimol, E.I.; Smeeth, L.; Guttmann, A.; Harron, K.; Moher, D.; Petersen, I.; Sørensen, H.T.; von Elm, E.; Langan, S.M.; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **2015**, *12*, e1001885. https://doi.org/10.1371/journal.pmed.1001885
29. Hang, T.; Innes, S.; Hope, J. Comparison of the Victorian emergency minimum dataset (VEMD) human intent coding descriptor to medical records. *Health Inf. Manag. J.* **2025**, *55*, 426–432. https://doi.org/10.1177/18333583251360616
30. Ashby, W.R. *An Introduction to Cybernetics*; Chapman & Hall: London, UK, 1956.
31. Hofstadter, D.R. *I Am a Strange Loop*; Basic Books: New York, NY, USA, 2007.
32. Luhmann, N. *Social Systems*; Stanford University Press: Stanford, CA, USA, 1995.
33. Zönnchen, B.; Dzhimova, V.; Socher, G. From intelligence to autopoiesis: Rethinking artificial intelligence through systems theory. *Front. Commun.* **2025**, *10*, 1585321. https://doi.org/10.3389/fcomm.2025.1585321.
34. Maturana, H.R.; Varela, F.J. *Autopoiesis and Cognition: The Realization of the Living*; D. Reidel: Dordrecht, The Netherlands, 1980.
35. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space. *Off. J. Eur. Union* **2025**. Available online: https://eur-lex.europa.eu/eli/reg/2025/327/oj (accessed on 8 November 2026).
36. Chen, J.; Chun, D.; Patel, M.; Chiang, E.; James, J. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 44. https://doi.org/10.1186/s12911-019-0793-0.

37. Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; McLachlan, S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 230–238. https://doi.org/10.1093/jamia/ocx079.
38. The MITRE Corporation. Synthea: Synthetic Patient Population Simulator, Master Branch, Commit aa0772fb (5 March 2026). Available online: https://github.com/synthetichealth/synthea (accessed on 30 March 2026).
39. Beale, T.; Heard, S. *openEHR Architecture Overview*; Release 1.0.2; openEHR Foundation: London, UK, 2008. Available online: https://specifications.openehr.org/releases/BASE/latest/architecture_overview.html (accessed on 30 March 2026).
40. Stummer, F.O. Enhanced Mycelium of Thought (EMoT): A bio-inspired hierarchical reasoning architecture with strategic dormancy and mnemonic encoding. *arXiv* **2026**, arXiv:2603.24065.
41. Kassenärztliche Bundesvereinigung. Gesundheitsdaten: Arztgruppen. 2024. Available online: https://gesundheitsdaten.kbv.de/cms/html/16744.php (accessed on 30 March 2026).
42. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed on 8 November 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.