Synthetic matched worlds for evaluating medical disengagement in type 2 diabetes: a mechanistic simulation benchmark of claims-based evaluation designs

Nov 6, 2026 (revised version, R1, v019i), Jérôme Vétillard - VP R&D Qualees

Abstract

Background. In claims data, a patient who disengages from chronic care is unobserved and his or her individual counterfactual is unobservable, so evaluations of vigilance programmes carry a bias of unknown size and sign.

Objective. The primary objective is methodological: to build a simulation benchmark in which truth is known by construction and to measure how specified claims-based evaluation designs recover specified causal contrasts of medical disengagement, with type 2 diabetes as a first use case. Detection with connected devices and a headroom analysis of the PREDICARE vigilance programme are secondary, worked examples under explicit assumptions.

Methods. predicare-synth is a mechanistic causal microsimulation of the cardio-reno-metabolic syndrome, not trained on patient data, with parameters posited from published statistics; calibration is pending. A daily trunk feeds simulated French claims (SNDS), connected devices (IoMT) and a patient app. A monthly trunk follows 60,000 people with type 2 diabetes for 20 years in matched worlds sharing common random numbers: continuous follow-up, status quo, re-engagement alone, an IoMT variant and the PREDICARE sequence (claims-based targeting, mobile-unit assessment, monitoring during disengagement). Estimands are defined explicitly. Deterministic results are replicated over 10 seeds with Monte Carlo standard errors (MCSE), and the claims designs were re-run under 14 alternative data-generating mechanisms. The performance gain of the IoMT registry is imposed as a prescribed curve: model training, label choice, temporal validation and transfer across territories are not simulated. Costs are in 2024 euros; the cost-utility analysis (payer and collective perspectives, 2.5% discounting, two probabilistic analyses of 300 draws) reports a headroom, the maximal fixed operating cost compatible with a reading threshold.

Results. Among persons meeting the claims silence criterion, the paired simulated contrast under the assumed model showed that disengagement raised 10-year mortality by 5.8 points (MCSE 0.1) while lowering undiscounted payer costs by 1,981 € (MCSE 76). Restricting the cohort to hospital reappearance biased the cost upwards by 3,087 € (MCSE 323; positive in every seed), intention-to-follow survival diluted the mortality effect (2.1 vs 5.3 points in the single-seed cohort emulation; 3.3 vs 5.8 points over 10 seeds) and as-treated analysis missed a delayed mortality. Across 14 posited, uncalibrated mechanisms, the upward bias of reappearance-based cohorts persisted (139 of 140 mechanism–seed pairs), but the intention-to-follow mortality bias reversed when disengagement was a marker of latent severity, and the sign of the economic headroom of re-engagement depended on the mechanism. As a secondary, conditional illustration, the headroom of the PREDICARE sequence (the maximal fixed operating cost compatible with 50 k€/QALY under posited, uncalibrated parameters) depended chiefly on the unmeasured attenuation of excess risk under monitoring and on attendance; it is not an estimate of what PREDICARE may cost. Under the posited selection on latent burden, comparing attendees with non-attendees overstated the mortality effect about sixfold; the defensible estimand is the intention-to-treat

effect of the offer policy among all eligible persons, which, under the simulated invitation budget, was diluted to −0.31 points of 10-year mortality. In an internal prediction experiment, simulated IoMT features raised the ROC-AUC for 30-day heart-failure decompensation from 0.57 to 0.70, with a positive predictive value of 5% at a 5% alert quota.

Conclusions. The main contribution is a simulation benchmark in which truth is known by construction, showing that, in the mechanisms tested, claims-based evaluation designs misstate the effect of medical disengagement: selecting on hospital reappearance overstated its cost under every mechanism, while the direction of the mortality bias depended on the causal mechanism posited. It also shows why the effect of the offer policy among all eligible persons, not the effect among participants, is the estimand a randomised evaluation can identify. The registry, prediction and cost-utility analyses are conditional illustrations; the benchmark does not show that PREDICARE works or what it may cost, and its uncalibrated results say nothing about French incidence, mortality or costs.

Keywords. medical disengagement; loss to follow-up; microsimulation; synthetic data; claims data; simulation study; estimands; cost-utility analysis; remote patient monitoring; type 2 diabetes; evaluation design

1. Introduction

Chronic disease now dominates health expenditure in France: about 13.7 million people (one in five) were covered by the long-term condition scheme (ALD) in 2021, accounting for about two thirds of health insurance spending [1]. These patients follow continuous trajectories, yet they are observed intermittently: a person with well-controlled type 2 diabetes is seen by a physician only a few times a year. When follow-up stops, nothing in the system raises a flag: the patient simply stops producing data.

We call this *medical disengagement* (in French, *déshérence*). Our working thesis is that it is not mainly a failure of professional attention but the expected outcome of a discrete observation regime applied to continuous risk. It is therefore an architectural property, which a vigilance programme can change only by changing what is observed, how often, and who acts on it.

PREDICARE is one such programme. Its first aim is to create a longitudinal registry of people in disengagement, through reinforced IoMT observation, in order to identify retrospectively the patterns that precede decompensations; the predictive and clinical value of those patterns remains to be established. Claims data would locate ruptures of follow-up and retrieve the events, a baseline mobile-unit assessment would characterise the initial clinical state, and connected devices would provide the repeated measurements between that assessment and any event — precisely what usual care data lack for patients who no longer consult. The intended sequence is to build the registry, link the series to clinically confirmed events, learn retrospectively, and validate on a later period and another territory, with alerts first computed in silent mode; only then would alerts be used to escalate patients to acute care. In the longer run, the programme is meant to preserve continuity of surveillance for people with long-term conditions despite declining medical density. This paper does not provide evidence for any of these steps. It addresses a prior question: whether the claims-based designs that would be used to evaluate such a programme can recover a known effect at all; and its registry is a prescribed performance curve, not a simulated learning process.

Three obstacles block the evaluation of such a programme on real data. First, the disengaged patient is invisible exactly while disengaged; studies typically identify them afterwards, when they reappear, frequently through hospital care. Second, the individual counterfactual is unobserved: the same patient, followed, cannot be observed. Third, the benefit of prevention is an avoided event, which is unobservable

at the individual level (the *prevention paradox*). Under these conditions an observational estimate cannot be checked against truth.

Synthetic data are usually proposed to protect privacy or to augment training sets. We use them for a different purpose: to build *worlds in which truth is known*, so that detection methods, targeting rules and evaluation designs can be scored before they are imposed on patients. The logic is the converse of target-trial emulation [2]: instead of emulating a trial in real data, we generate data around a known causal mechanism and check whether the designs applied to those data recover it, as in simulation studies that evaluate statistical methods [3].

The main contribution of this paper is a simulation benchmark in which truth is known by construction, showing that, in the mechanisms tested, claims-based evaluation designs misstate the effect of medical disengagement: selecting on hospital reappearance overstated its cost under every mechanism, while the direction of the mortality bias depended on the causal mechanism posited. Its primary objective is therefore methodological: to measure how specified claims-based designs recover specified causal contrasts, under the posited mechanism and under alternative mechanisms in which disengagement is partly or wholly a marker of severity. The prediction experiment with connected devices and the cost-utility analyses are secondary; they show what the benchmark can be used for, and what it cannot establish.

This paper makes five contributions:

1. a mechanistic generator of cardio-reno-metabolic trajectories with SNDS-like, IoMT and app observation heads, in which every parameter carries an explicit validity domain and epistemic status;

2. a matched-worlds design based on common random numbers that yields simulated counterfactual outcomes known by construction, so that the mean effects of disengagement and of re-engagement are measured without comparison bias, conditionally on the posited model;

3. a claims-only targeting script, transposable to the real SNDS, whose precision, recall and estimation bias are measured against the simulator's ground truth (not against real-world disengagement);

4. evidence that, in the configurations tested, the claims-based observational designs tested (cohorts defined by reappearance, intention-to-follow survival and as-treated attribution) each miss or overstate the effect of disengagement; that, within a class of 14 posited data-generating mechanisms, the upward bias of reappearance-based cohorts persists while the direction of the mortality bias does not; and a cost-utility framing that replaces the question "how much will it save?" by "how much may it cost?", reported as a headroom in three perspectives and under probabilistic sensitivity analysis;

5. a model of the PREDICARE sequence (world PE) (claims-based targeting, mobile-unit assessment at T0 and IoMT monitoring that starts during disengagement) showing that, under a posited selective attendance, the naive attendee-versus-non-attendee contrast overstates the effect about sixfold, and that the evaluation must randomise the offer.

2. Methods

2.1 Study design

The benchmark separates three things that real data merge: the mechanism (a causal model), the observation of that mechanism (claims, devices, app), and the analyses. Analyses read only observations; ground truth is stored apart and read only by the evaluation step (Figure 1).

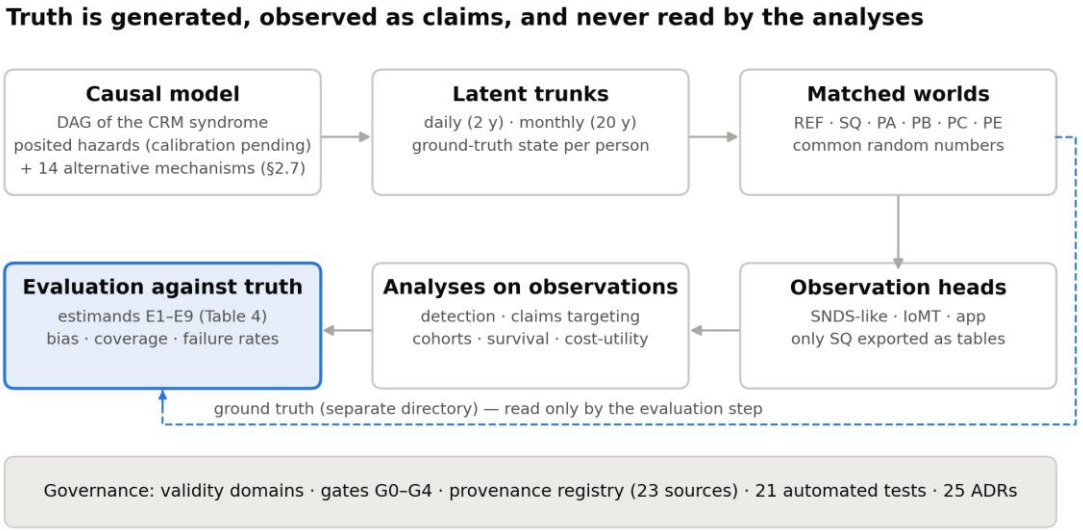

Figure 1 - Benchmark pipeline: six stages and one truth path. Hazards are posited, not calibrated; counterfactuals are known under the assumed model only.

Each experiment compares arms that differ in one factor only (a feature set, a design, a queue order or a world) at identical latent truth, population and data split. Results are therefore relative by construction. The study is organised following the ADEMP structure for simulation studies (aims, data-generating mechanisms, estimands, methods, performance measures) [3]: estimands are defined in §2.6 (Table 4), alternative data-generating mechanisms in §2.7, and every deterministic long-term result is replicated over 10 seeds (§2.4).

2.2 The generator: a mechanistic causal microsimulation, not a trained model

No patient data were used to train the generator. predicare-synth is not a deep generative model (GAN, VAE, diffusion or language model) fitted to real records. It is a stochastic microsimulation whose structure is a causal model written by hand and whose parameters are posited from published aggregates (formal calibration is pending, see §2.5 and Supplementary Table S2). This choice is deliberate. A generator trained on real data reproduces the observation regime of those data (including their blind spots) and cannot produce the very thing that is missing: longitudinal ambulatory trajectories of patients who have stopped being observed.

Causal structure. A directed acyclic graph (DAG) links the conditions of the cardio-reno-metabolic syndrome, following the American Heart Association CKM framework [4]: obesity → type 2 diabetes, hypertension, dyslipidaemia → chronic kidney disease staged by eGFR (3a, 3b, 4, 5) [5] → heart failure, with myocardial infarction and stroke as events. COPD is kept as an out-of-graph control condition. Edges carry multiplicative hazard ratios of the order reported in the literature (e.g. diabetes multiplies the hazard

of CKD by ~3.5). The graph is acyclic by construction; the cardio-renal loop is broken in the majority metabolic → renal → cardiac direction and documented as a simplification.

Two latent trunks, one parameter source (Table 1).

Table 1 - The two latent trunks: time step, horizon, population and purpose.

Trunk	Time step	Horizon	Population	Purpose
Daily	1 day	2 years	40,000–100,000 adults, CRM + COPD	short-horizon detection (≤ 30 days), IoMT signals
Monthly	1 month	20 years (1 + 10 + 10)	60,000 people with type 2 diabetes	long-term outcomes, costs, QALYs, cohort emulation

Both trunks read the same hazard parameters from the same configuration files, so that a change in the causal model propagates to both.

Latent state. Each simulated person carries a ground-truth state that no rule or model ever reads: conditions with dated onset, CKD stage, daily physiology in the daily trunk (weight, blood pressure, glycaemic index and CGM-derived time-in-range, resting heart rate, activity, sleep, SpO_2, eGFR, potassium), a three-domain microvascular burden in the monthly trunk (retinopathy 0–2, albuminuria A1–A3, diabetic foot grade 0–2), treatment state, and disengagement episodes.

Disengagement. Rupture is drawn as a hazard while followed (~4.7% per year) and is recurrent: spontaneous return, reconnection after a hospital stay (60%), and later re-rupture. Four months into a rupture, decompensation hazards are multiplied by 3.5, cardiovascular hazards by 1.5, and progression of CKD and microvascular damage is accelerated (×1.8 to ×2.5). A partial form keeps pharmacy dispensing alive for 2–6 months.

What "disengagement" means in the model. A rupture combines three mechanisms that a claims study cannot separate: loss of observed contact (no GP visit, no biology), interruption of treatment (no dispensing, except during a partial rupture with residual pharmacy), and an excess clinical risk that starts four months after the onset of rupture. *Medical disengagement* refers to the latent rupture; *observable silence* refers to its claims trace (estimand E8, §2.6). In the main analyses the whole excess risk is caused by the rupture; §2.7 relaxes this assumption.

Microvascular hierarchy. The burden defines three strata (low, intermediate, high). A posited gradient (cardiovascular ×1.0/1.4/2.0; diabetic decompensation ×1.0/1.2/1.5; mortality ×1.0/1.2/1.5) follows the direction reported by population cohorts in which cardiovascular and amputation risk rise with the number of microvascular complications [6,7].

Therapeutic escalation (monthly trunk). Metformin is the backbone; SGLT2 inhibitors, GLP-1 receptor agonists and basal insulin are initiated while followed, interrupted during rupture, and resumed on return, respecting French availability (GLP-1 RA from 2009, SGLT2i reimbursed from 2020). Their effects are pooled hazard ratios from meta-analyses of cardiovascular outcome trials [8,9] (Table 2).

Table 2 - Treatment effects applied while the patient is on treatment.

Outcome	GLP-1 RA (8 CVOTs)	SGLT2i (6 CVOTs)
MACE (applied to MI, stroke)	0.86 (MI 0.91, stroke 0.84)	0.90
Hospitalisation for heart failure	0.90	0.68
Composite kidney outcome (CKD incidence and progression)	0.83	0.62
New macroalbuminuria	0.74	—
Background mortality (residual, posed)	0.95	0.95

Mortality and dialysis. Background mortality follows a Gompertz law multiplied by heart failure, advanced CKD, dialysis, microvascular stratum and rupture; acute events carry their own case fatality. Stage-5 CKD leads to dialysis, unplanned when started during rupture.

Parameter sources. Every external figure enters through a provenance registry (23 sources in v0.17: identifier, producer, URL, licence, intended use, status, SHA-256 fingerprint once downloaded). Sources include the Health Data Hub synthetic SNDS "top diabète" dataset [10] (used for schema alignment only, not for epidemiology), the Cnam disease-expenditure dataset (2024 vintage, downloaded and fingerprinted) [11], the INSEE consumer price index [12], the REIN registry [13], NHANES, KDIGO [5], the AHA CKM statement [4], heart-failure telemonitoring studies [14,15], UKPDS 62 [16] and the CVOT meta-analyses [8,9]. A generation run refuses to start if the registry or the governance frame is non-compliant.

2.3 Observation models, interventions and matched worlds

Observation heads. The latent state is projected onto three channels with their own sampling and missingness, never onto the truth itself (Table 3).

Table 3 - Observation heads: what each channel records, its cadence and its missingness.

Channel	What it records	Cadence	Missingness
SNDS-like claims	dated acts without results: dispensing (ATC), GP and specialist visits, biology acts, hospital stays (PMSI motive, length, entry by emergency), emergency visits, long-term condition (ALD) dates, death	event-driven; a followed diabetic ≈ 4 GP visits/year	structural: silence after rupture; deaths flagged at 98%
IoMT kit	weight, blood pressure, CGM, resting HR, activity, sleep, SpO_2, plantar temperature	daily	active devices stop ~15 days before administrative rupture; passive ones ~30 days after; optional MNAR coupling before decompensation
Patient app	declared symptoms, structured monthly check-in	spontaneous or solicited	non-response is a recorded state, distinct from absence of symptom

The SNDS head was aligned with the schema of the Health Data Hub synthetic SNDS release (ER_PRS_F, ER_PHA_F, IR_IMB_R, PMSI tables). Nomenclatures are symbolic labels with candidate codes flagged for confirmation.

Intervention model. Each month, a vigilance queue lists patients whose *observable* silence exceeds the targeting rule (> 90 days without dispensing and > 180 days without GP visit). Under a capacity budget (0.2% of the living population per month in the long-term trunk, 2–5% in short-horizon experiments), the queue is ordered either by longest silence (policy A) or by observable microvascular burden (policy B). A contact succeeds with probability 0.60 and leads to return with probability 0.50; return extinguishes the excess risk after one month and resumes treatments. Each alert costs 40 €. The return probability is the least documented parameter of the benchmark and is treated as a policy variable.

Matched worlds with common random numbers. Following the common-random-numbers principle of disease simulation [17], each month consumes a fixed array of uniform draws whose columns have a fixed meaning (rupture, return, each event type, death, progression, each observation, contact). All worlds read the same array, and an event occurs when the draw falls below the world's hazard. Two worlds therefore differ only where their hazards differ, and are identical for a patient until his or her first rupture, a property enforced by automated tests. Four core worlds are run jointly (v0.15 adds the IoMT worlds PC/PD and the PREDICARE sequence, world PE, described below):

- **REF** — continuous follow-up (rupture hazard set to zero);
- **SQ** — status quo (ruptures, spontaneous returns, reconnection after hospital);
- **PA** — re-engagement alone, queue ordered by longest silence (policy A);
- **PB** — re-engagement alone, queue ordered by observable microvascular burden (policy B).

Averaged over patients, SQ − REF measures the effect of disengagement and PA − SQ the effect of re-engagement (paired simulated population contrasts, §2.6), without comparison bias and with reduced Monte Carlo variance. These simulated counterfactual outcomes are known by construction, not observed: they hold only conditionally on the posited mechanism, and differences for an individual patient additionally depend on the chosen coupling of random draws, a variance-reduction device, not an identification of individual counterfactuals. We therefore report population and stratum averages only. Only the SQ world is exported to SNDS-like tables; the others remain in a separate ground-truth directory.

IoMT variant after re-engagement (worlds PC/PD, v0.15). In PA and PB, a re-engaged patient returns to standard follow-up. PC and PD test a variant in which the re-engaged patient is additionally equipped with an IoMT kit and becomes better followed than a standard patient with diabetes, and the daily measurements feed a longitudinal registry whose detection performance is assumed to improve with the episodes it accumulates. Worlds PC (queue A) and PD (queue B; implemented and tested, not reported here since queue choice is examined in §3.3) implement this on a dedicated random stream, so that SQ and PA are unchanged. On return obtained by contact, the patient accepts the kit with probability 0.80 and abandons it at 2% per month; the signal stops during rupture, and its loss makes the patient a queue candidate the following month, ahead of claims-based silences. A heart-failure, acute-kidney-injury or diabetic decompensation on its way in an equipped, followed patient is detected with sensitivity s(t) and then avoided with probability 0.40. The registry's performance gain is prescribed, not simulated (no learning algorithm is run): $s(t) = s_{min} + (s_{max} - s_{min}) \times R/(R + R½)$, where R is the cumulative number of decompensations seen in equipped patients, $s_{min} = 0.05$ is chance at a 5% alert budget, and s_{max} (0.24 for heart failure, 0.30 for kidney injury, 0.20 for diabetic decompensation) and the avoidability are

carried over from the daily trunk, a parametric coupling of the two trunks. The heart-failure ceiling is the operational recall measured in the internal prediction experiment (0.24 at a 5% per-date quota, §3.1); the earlier value (0.39, from a pooled top-5% budget in the v0.10 daily trunk) is retained as an optimistic sensitivity case. This bridge is a coupling hypothesis, not an empirical validation, and so is the avoidability of a detected event (0.40), carried over from the daily trunk. Its differences are handled as follows. (i) Unit of the quota: in the daily trunk, index dates are 30 days apart and the quota flags 5% of equipped patients at each date; in the monthly trunk, the quota is 5% of equipped persons per month, so the nominal alert rate is the same, and only the cost of alerts is tied to it. (ii) Repeated alerts and deduplication: in the daily trunk a patient can be flagged at successive dates and each flag counts as an alert; the monthly trunk does not simulate alerts one by one but draws, for each incoming decompensation of an equipped person, whether it is detected with probability s(t), so repeated alerts, their deduplication and the number of windows preceding an event are not represented. (iii) Episode definition: in the daily trunk a positive is a hospitalisation for acute heart failure within 30 days, counted at one landmark at most; in the monthly trunk it is a decompensation occurring in the month. (iv) Event rate: equipped persons in PE are monitored during disengagement and decompensate more often than the equipped heart-failure patients of the daily trunk, which changes recall at a fixed quota in a direction we cannot sign. None of these differences is corrected; we bound their consequence by ceilings of 0.12 and 0.06 (§3.6). The registry does not learn in this paper: no learning algorithm is run in the long-term worlds, and its performance gain is a prescribed curve. Equipment costs 25 € per equipped patient per month in the code default (110 € in the reported results of world PE; §2.4) plus 20 € per alert. In the long-term worlds the alert budget is a monthly quota (5% of equipped persons per month), not a score threshold. A real registry would have to define positive episodes (a hospital decompensation within 30 days of an index date), negatives, a retraining schedule, out-of-period validation and transferability across patients and territories; none of this is simulated, and the curve is a plausible assumption of performance gain, not the result of a learning experiment. Variants restrict equipment to patients with observable heart failure or CKD, pre-seed the registry, or route signal losses to a dedicated telemonitoring queue outside the re-engagement budget.

PREDICARE sequence: mobile-unit assessment and monitoring during disengagement (world PE). PC adds IoMT to re-engagement; the intended programme starts earlier. World PE chains five steps: (1) claims-based targeting of silent patients under a monthly invitation budget; (2) invitation, reachability (0.60) and attendance at a mobile health unit ("médicobus", 0.50 if reached), modulated by the latent microvascular burden (×1.15, ×1.0, ×0.75), a posited selection; (3) baseline clinical and biological assessment at T0 (180 €), which makes CKD and heart failure known; (4) equipment (0.80) and monitoring that runs during disengagement, with its own attrition (4% per month while disengaged), whether or not conventional follow-up resumes (0.35); (5) alert and response: detection with the prescribed registry sensitivity, avoidance (0.40), and reconnection to care after a true alert (0.50). Supplementary Figure S1 shows the pathway with its cost entry points (C1–C4) and health-effect entry points (H1–H5). Two policy variables are undocumented and swept: the share of the disengagement excess risk erased by remote monitoring (0 to 1) and the renewal of treatments by the telemonitoring team. Evaluation keeps the denominator of all invited patients and separates the effect of the offer (intention to treat, matched truth), the effect among attendees (a principal stratum, estimated by the complier average causal effect, CACE = ITT / attendance, valid under the exclusion restriction and one-sided non-compliance [18]), and the naive attendee-versus-non-attendee contrast.

T0 decisions and clinical capacity (v0.16). Two mechanisms were added to PE, with defaults intended to reproduce v0.15 (the automated test checks only that a non-binding alert capacity leaves world PE unchanged). (i) The T0 assessment measures CKD, heart failure and a high microvascular stratum with posited sensitivities and false-positive rate; under a stratified policy only attendees classified at high risk are equipped. (ii) Alerts are handled by a team whose capacity is capped (alerts reviewed per 100 equipped

patients per month, against an alert budget of 5), optionally prioritised by T0 risk; only handled alerts are costed, and avoidability is scaled by g(d)/g(2) with g(d) = max(0, 1 − d/13), d being the response delay in days and 13 days the median lead time of the daily trunk.

2.4 Analyses

Short-horizon detection (machine learning; secondary). Prediction of a hospital decompensation within 30 days used a landmark design: monthly index dates, features strictly observable at the index (claims cadences and history; IoMT slopes and levels over 7–30 days; app declarations; biology values when present), outcome in (t, t + 30 days]. The learner was the same in every arm, scikit-learn HistGradientBoostingClassifier (300 iterations, native handling of missing values) [19], with a train/test split by patient frozen before any filtering. Only the feature set changed between arms. Historical campaigns (v0.2–v0.11, 3 to 5 seeds; Supplementary Table S9) varied prodrome length, noise and equipment rate, ablated port-count versus physiological values, and coupled equipment to disengagement (social gradient). For heart failure, the experiment was re-run as an internal simulation experiment, explicitly not a validation: 10 seeds; patients disjoint between training (50%), calibration (20%) and test (30%); training on index dates up to day 420 and testing after; alert budget applied as a per-date quota; cluster-bootstrap intervals over patients within seeds (400 replicates). Metrics were ROC-AUC, PR-AUC, recall and positive predictive value at the quota, Brier score and calibration slope.

Evaluation designs. On matched worlds we compared the true avoided events (ITT, by world difference) with estimators a real programme could compute: alerted-returned versus alerted-not-returned, pre/post among returners, an emulated individual RCT, a stepped-wedge and an inverse-probability-weighted (IPTW, logistic propensity) adjustment of the naive contrast. Each estimator was assessed over 10 seeds (daily trunk, 40,000 persons per seed) by its ratio to the truth, the coverage of its 95% bootstrap interval (B = 300) and its failure rates (wrong sign; relative bias above 50%) (Supplementary Table S8).

Cohort emulation from claims. A targeting script reading only the SNDS-like tables defined:

- *population*: long-term condition 8 (diabetes) opened before 2007;
- *silence*: > 90 days without dispensing and > 180 days without GP visit, no known death;
- *index*: last contact before the first silence observed in 2007–2016, guaranteeing ≥ 10 years of follow-up;
- *competing exits*: hospital reappearance, emergency only, return to primary care, death, persistent silence, at 5 and 10 years;
- *comparator*: 1:1 exact matching with replacement at the index month among patients alive, in contact within 6 months and never silent before. Level-1 keys were 5-year age band, sex, diabetes duration class (< 5, 5–9, ≥ 10 years since ALD opening), treated hypertension and dyslipidaemia (12 months), therapeutic intensity (metformin only / one add-on / insulin or ≥ 2 classes), tertile of reimbursed cost over the previous 12 months, observable microvascular stratum, CKD and heart-failure ALDs. Levels 2 and 3 relaxed the keys hierarchically. Balance was assessed by standardised mean differences (SMD) [20].

Identical event variables were built for both cohorts: date and delay of first hospitalisation, first decompensation (heart failure, acute kidney injury on CKD, diabetic decompensation), first emergency visit and death, with censoring at death or 31 December 2026; life-years and cost per life-year in each

window; cost split into pharmacy, ambulatory care, dialysis, hospital and emergency; recurrent silence episodes.

Survival. Kaplan–Meier curves [21] (intention to follow); cumulative incidence by the Aalen–Johansen estimator [22] with death as a competing risk; as-treated curves censoring each patient when he or she changes state (return to care for the disengaged, first later silence for the followed); and incidence rates per person-year by time-varying exposure state.

Economic evaluation. Base perspective: the French statutory insurer (the collective perspective of the HAS reference case is described below); all costs are expressed in 2024 euros (Supplementary Table S1-C). Items taken from the Cnam disease-expenditure dataset [11] are read from its 2024 vintage (outpatient care and hospital stays; cash benefits excluded): diabetes 2,296 € per year, of which a posited ambulatory fraction of 60% is applied to followed months, and chronic dialysis 67,538 € per year. Acute events (heart failure 5,824 €, acute coronary syndrome 8,915 €, stroke 14,209 €) are valued by the hospital component of the episode year, not by the annual cost of that year, which would double-count outpatient care already valued monthly; the annual total is kept as an upper-bound sensitivity analysis. Posited items are declared at 2024 prices and flagged. Variable programme costs (contact or invitation 40 €, T0 assessment 180 €, equipment and platform 110 € per equipped month, alert handling 20 €) are modelled explicitly (the equipment cost includes about 86 € per month for a continuous glucose sensor at the French reimbursement base; the code default of 25 € is corrected exactly, the cost having no feedback on trajectories; Supplementary Table S14); the fixed operating cost is not. Utilities: baseline 0.785 with additive decrements from UKPDS 62 [16] (MI −0.055, heart failure −0.108, stroke −0.164, amputation −0.280) and, for end-stage renal disease (−0.263) and dialysis (−0.305), from values reported by McEwan et al. and Brändle et al. as tabulated in a systematic review of diabetes economic evaluations [23] (a review of insulin models in type 1 diabetes: these are second-hand values, and the primary type 2 sources must be checked and cited directly, or replaced, at G2); the baseline lies within the range reviewed by Beaudet et al. [24]. Costs and outcomes were discounted at 2.5% per year, as recommended by the HAS for horizons under 30 years [25]. Because the fixed operating cost F of each strategy (re-engagement alone or the PREDICARE sequence) is unknown, it was treated as a parameter (€ per covered patient per year, applied to discounted covered patient-years, about 11 per person with diabetes over 20 years): ICER(F) = (ΔC + F × covered patient-years) / ΔQALY, and we report the *headroom*, the maximal F compatible with reading thresholds of 30, 50 and 100 k€/QALY (France has no official threshold). The headroom is a threshold quantity, not a budget: it is expressed per person with type 2 diabetes in the covered territory (estimand E4) and must cover only resources that do not vary with volume (mobile unit, base staffing, platform licence, coordination; Supplementary Table S1-C). The economic results are therefore a worked example of a headroom analysis, conditional on declared hypothetical effects, not an economic estimate of PREDICARE. Univariate scenario analyses covered 13 scenarios, each run over 10 seeds. Because the HAS reference case is a collective perspective including informal care [25] (productivity losses only in a supplementary analysis), v0.15 adds a valuation layer without feedback: survivors of stroke, amputation, dialysis start or heart-failure decompensation become dependent at home or institutionalised with posited probabilities, alongside age-related loss of autonomy; dependency is valued with DREES 2019 unit costs [26] converted to 2024 euros with the INSEE consumer price index (factor 1.1496) [12] (555 € per month at home; 2,155 € accommodation plus 586 € dependency per month in nursing homes) and posited informal-care hours (to be confronted with the CARE surveys [27]) at replacement cost. A probabilistic sensitivity analysis drew 23 parameters jointly, following ISPOR–SMDM recommendations on uncertainty analysis [28], from declared distributions (lognormal, beta, uniform) centred on base values (disengagement mechanism, re-engagement, costs, utilities and IoMT) and re-simulated the matched worlds for each of 300 draws (20,000 patients per draw); a second analysis on world PE added seven distributions specific to the sequence (Table S3).

Monte Carlo error and value of information. Deterministic long-term results use 10 seeds (42–51) of 60,000 people (30,000 for the collective-perspective layer) and are reported as mean ± Monte Carlo standard error (MCSE: between-seed standard deviation divided by √10), with [min; max] over seeds; paired contrasts are computed seed by seed. Differences smaller than about two MCSE are not interpreted. In the probabilistic analyses each draw simulates 20,000 people, so the raw distributions combine parameter (second-order) and Monte Carlo (first-order) uncertainty. The first-order variance, estimated within draws, was subtracted from the between-draw variance (analysis of variance [29]); it accounted for 5–12% of the variance of the reported quantities, so the raw intervals slightly overstate parameter uncertainty. For the PREDICARE sequence, the incremental net benefit of a draw was 50,000 × ΔQALY − ΔC − F × covered person-years; the expected value of perfect information (EVPI) and of partial perfect information (EVPPI) were estimated from the 300 draws by nonparametric regression of this net benefit on the parameters [30] (cubic polynomial for each parameter, additive fit for groups), with bootstrap standard errors over draws; values not distinguishable from their standard error are not interpreted.

2.5 Governance and quality control

The benchmark applies to itself the governance it argues for: no claim may travel beyond the validity domain of the component that produced it. These controls discipline development and make assumptions inspectable; they are not a substitute for clinical validation or for independent review of the causal model [31].

Validity domains. Each component (ontology, DAG, each observation head, each emulator, valuation, utilities, provenance registry, 21 components) declares in a machine-readable file what it allows one to assert, what it forbids, and its current epistemic status. Every component is at status *hypothesis* except the provenance registry (*convention*). For example, the long-term trunk allows statements about the *shape* and *sign* of differences between matched worlds and forbids any real incidence, mortality or euro amount.

Decision gates. Status can change only through a gate, never by drafting: G0 adoption of the frame at kick-off; G1 clinical workshop (hypothesis → convention); G2 calibration against a registered source (→ calibrated); G3 the demonstrator's go/no-go; G4 confrontation with real data (→ corroborated). Each gate names an authority — some not delegable to the technical team — and what passing it does *not* authorise. An automated check refuses illegal transitions; during development it caught and blocked one.

Provenance. No external datum enters without a registry entry; a downloaded source must carry its SHA-256 fingerprint, and every generation run writes a manifest with the fingerprints of the ontology, physiology, governance and registry files.

Determinism and invariance. Random streams are isolated per patient × component × sub-stream through NumPy SeedSequence, so that changing an equipment policy does not alter latent trajectories (an early bug of this kind was detected and corrected). A given seed reproduces the data byte for byte.

Automated tests (21; Supplementary Table S13). A recette pipeline checks each seam of the application chain (targeting → individual twin → drift detection → queue → cockpit), including a deliberately failing negative test: when a person outside the observable silent set is injected into the targeted list, the chain must fail. Further tests verify determinism; identity of continuous follow-up and status quo before a person's first rupture; identity of REF, PA and PB with SQ for persons never in rupture; that world PC equals PA without equipment, and that adding world PC or activating a structural mechanism leaves the other worlds unchanged (isolated random streams); that a non-binding alert capacity leaves world PE unchanged; strict separation between the claims tables and the truth directory (the targeting script refuses

to run on the latter); and minimum precision (> 0.85) and recall (> 0.80) of claims-based targeting against the simulator's rupture definition. That v0.18 reproduces v0.17 byte for byte under its default structure was verified separately on run outputs. None of these tests checks the clinical correctness of the mechanisms.

Decision records. Twenty-five architecture decision records (ADR-001 to ADR-025) document every structural choice, its rejected alternatives and its consequences.

Adversarial review. A written contradiction file lists the ten strongest objections to the programme (circularity, thresholds, paternalism, false positives, social gradient, unobservable avoidance…), with the current position, the weakness and the evidence that would settle each. A critical review of the benchmark (13 biases) and an external epidemiological review of the cohort workbook (asymmetric outcome variables, missing confounders) led to corrections reported here.

Model validation. Following ISPOR–SMDM Task Force 7 [31]:

- *face validity,* the causal structure and parameters were examined by the author and by the adversarial review above, not yet by an independent clinical panel;
- *verification,* the automated tests and recette pipeline above, including invariance checks between versions and worlds (Supplementary Table S13);
- *cross-validity,* no comparison with an established type 2 diabetes model (for example the UKPDS Outcomes Model) has been made;
- *external and predictive validity,* none. Calibration against French incidence and mortality is planned at gate G2; Supplementary Table S2 lists the simulated aggregate rates, several of which (notably dialysis starts and myocardial infarction) appear high against French orders of magnitude. The benchmark is therefore validated, at most, for its stated use (testing whether methods recover a known truth) and not for predicting French outcomes.

2.6 Estimands

Every quantity reported in this paper is one of the estimands in Table 4, or an estimator of one of them. Three kinds are distinguished throughout. *Paired simulated population contrasts* compare the same simulated population under two strategies; the two worlds share common random numbers, which reduces Monte Carlo variance, and the contrasts are averages over a population, reported only as such. *Model-defined subgroup contrasts* are contrasts in subgroups defined by an outcome in one world, for example "patients who reappear at hospital in the status quo"; these are principal strata [32], whose values depend on the posited model and on the chosen coupling of random draws, and which are not identifiable in real data. *Observational estimators* are what a study restricted to claims could compute; their bias is the difference between the estimator and the corresponding paired contrast.

What "removing disengagement" means. The contrast SQ − REF (E1) is the effect of a hypothetical intervention that would prevent every rupture, keeping each person in the observation, treatment and risk regime of a followed patient. It bundles the three components of a rupture (§2.2) and does not correspond to a single health policy: in reality disengagement mixes interrupted access, patient choice, silent illness, relocation, follow-up outside the reimbursed system and treatment interruption, which call for different responses. The structural experiment varies some of these sources (care outside the system, partial ruptures, reverse causation; §2.7), not all. The policy-relevant contrasts are those of an intervention that can actually be offered: re-engagement (PA − SQ) and the PREDICARE offer (E5).

Time is counted in months of simulation. Month 0 is January 2006, follow-up starts at month 12, and the entry window for indexing runs over months 12–131, so that every index is followed for at least 120 months and there is no censoring before 10 years except by death. Death is part of each outcome: costs and QALYs accrue only while the person is alive. Returns to care, re-ruptures, treatment escalation and treatment interruption all occur within the horizon and are not censoring events in the paired contrasts; in the as-treated survival analysis (Table 7), a change of state is a censoring event by design of that analysis.

Table 4 - Estimands: population and denominator, time zero, contrast, outcome and status.

#	Estimand (where reported)	Population, denominator and time zero	Contrast, outcome and summary measure	Status
E1	Effect of disengagement among the claims-identified silent (Table 5; truth row of Table 6; Figure 3)	Persons with T2D in the status-quo world who first meet the observable silence criterion (> 90 days without dispensing and > 180 days without GP visit) during months 12–131: all such persons (about 18,100 per 60,000), whatever their later course. Time zero: last observed contact before silence	SQ vs REF for the same simulated persons; death by 5 and 10 years, cumulative payer cost by component (undiscounted), life-years in the window; difference in means	Paired contrast, but on a population selected on an event (silence) that occurs only in SQ: model-defined, conditional on the coupling
E2	Cost and mortality in outcome-defined subgroups (Table 9; Supplementary Table S12; Figure 2)	Subsets of E1 defined by the first exit observed in SQ and/or by the latent microvascular stratum at time zero	As E1	Model-defined principal stratum; not identifiable in real data. Only rows by claims-observable stratum describe subgroups a real study could form
E3	Claims-based estimators of E1 (Table 6, rows 2–4; Table 7)	(a) All silent persons vs 1:1 exactly matched followed persons (with replacement, at the index month); (b) the subset of (a) that reappeared at hospital vs the same comparators. Time zero: index month	Observed silent vs observed followed, both in SQ; outcomes as E1, plus Kaplan–Meier, Aalen–Johansen and as-treated survival and rate ratios by time-varying state	Estimators; bias = estimator − E1
E4	Population cost-utility of a strategy (Tables 8, 10, 11; Figures 5 and 6;	All simulated persons with T2D alive at month 12 (60,000 per seed; 20,000 per draw): every	Strategy world (PA, PB, PC, PE, REF) vs SQ, each a dynamic monthly policy;	Paired population contrast. The headroom is a threshold quantity, not a budget, and

	Supplementary Tables S11–S12; both probabilistic analyses)	person with T2D in the covered territory, not the targeted, invited or equipped persons. Time zero: month 12	discounted costs and QALYs at 10 and 20 years; ΔC per person, ΔQALY per 1,000, headroom F_max(λ) = (λ·ΔQALY − ΔC) / covered person-years (about 11 discounted person-years per person over 20 years)	excludes nothing that the model already counts as a variable cost (Supplementary Table S1-C)
E5	Effect of the offer (intention to treat) among all eligible persons, PREDICARE sequence (§3.5; Figure 6, right)	All persons who first meet the claims targeting criterion (observable silence) during months 12–131, whether or not they are invited (budget-limited queue; re-invitation possible 6 months after a failed contact): the population of E1, about 18,100 per 60,000. Time zero: month the criterion is first met. Under cluster randomisation, the same rule applied in both arms before allocation defines the denominator	PE vs SQ for the same persons (trajectories identical up to the first invitation, which requires silence); death by 10 years and 10-year payer cost; risk difference, € per eligible person	Paired contrast. Effect of the simulated policy, budget and queue included (not of an invitation received by each eligible person). Primary estimand for the demonstrator, identifiable by randomising clusters before targeting starts, each with its own capacity (§3.5). The effect among persons actually invited (time zero: first invitation) is secondary: that population is selected by the programme's queue and budget
E6	Effect among attendees (Figure 6, right)	Eligible persons who attend at least once within the horizon after repeated invitations (18.3% of E5 in the reference scenario; 67.1% of those invited)	PE vs SQ, as E5; CACE = ITT / first-stage effect on attendance	Principal stratum (compliers), identified by the CACE under one-sided non-compliance (holds by design) and the exclusion restriction (holds by construction; must be argued in the field). Secondary
E7	Naive attendee vs non-attendee contrast (Figure 6, right)	Attendees vs non-attendees among the invited, both in PE (time zero: first invitation)	Difference in observed 10-year mortality	Estimator, biased by the posited selection; shown only to quantify that bias
E8	Accuracy of claims-based targeting (Supplementary Table S9-D)	Living persons with an observable diabetes signal, assessed each month	Rule output vs simulator rupture state; precision, recall	Agreement with the simulator's own rupture definition, by a rule built in the same universe; not

				a performance estimate for the real SNDS
E9	Short-horizon discrimination of the IoMT detector (§3.1; Supplementary Table S9)	Equipped persons with claims-known heart failure at landmark dates every 30 days; test patients disjoint from training, index dates after day 420	Claims only vs claims + IoMT (same learner); ≥ 1 hospitalisation for acute heart failure in (t, t + 30 days]; ROC-AUC, PR-AUC, calibration, recall and PPV at a per-date quota	Internal simulation experiment on an injected prodrome; not model validation

Units and interference. Unless stated otherwise, a euro amount in Tables 8, 10 and 11 is "per person with type 2 diabetes in the covered territory" (E4). When the text translates the headroom into amounts per invited person, per attendee or per equipped patient-year (§3.5), it redistributes the same discounted total over a smaller denominator; these translations are labelled as such and never replace E4. Persons interact only through shared, capacity-limited resources (monthly budgets, alert-handling capacity and the registry curve), so E4 and E5 are defined for the policy as a whole, under this interference; §3.5 states which effect each possible trial design would estimate. The expression "same patients" is reserved for paired contrasts and always means "under the assumed model and coupling".

2.7 Structural experiment: alternative data-generating mechanisms

The main analyses explore parameter values within a single causal structure, in which a rupture causes the whole excess risk. To test whether the design biases are properties of claims-based evaluation or of that structure, we re-ran the claims designs under 14 data-generating mechanisms that change *who causes what* (v0.18; Supplementary Table S7). Left at its defaults, the structural block of v0.18 reproduces v0.17 byte for byte (verified on run outputs; an automated test checks that the structural stream leaves world REF unchanged).

- **Causal reference (M0).** The mechanism of §2.2.
- **Disengagement as a marker of latent severity (M1–M3).** A latent frailty Z (lognormal, mean 1, standard deviation of log 0.6), never observed, multiplies the rupture hazard and, in every world, the hazards of events and background mortality. Only a fraction θ of the rupture excess risk is causal: θ = 0.5 (M1) or θ = 0 (M2). In M3, θ = 0 and treatment also continues during rupture, so that disengagement causes nothing clinically.
- **Confounding by deprivation or access (M4).** Full causal effect, plus a shared frailty acting on rupture and, more weakly, on events.
- **Reverse causation by acute illness (M5).** The rupture hazard is multiplied by 4 in the 3 months after any hospital stay.
- **Partial and intermittent disengagement (M6).** Residual pharmacy in 80% of ruptures, for 6–12 months, and fast spontaneous returns.
- **Care outside the observable system (M7).** 30% of claims silences are not clinical ruptures: care and treatment continue, unrecorded and unpaid by the insurer.

- **Absent, reduced or delayed reconnection (M8, M8b).** No or reduced reconnection after a hospital stay; excess risk extinguished over 6 months after return.
- **Degrees of treatment discontinuation (M9, M9b).** Half, or none, of the treatment effect is lost during rupture.
- **Magnitude and timing of the effect (M10, M11).** Excess risk halved and starting in the first month; or delayed to month 8.

For each mechanism and each of 10 seeds (60,000 persons), we computed the paired truth (E1, 10 years) and the claims estimators (E3): all silent persons vs matched followed persons (cost); persons who reappeared at hospital vs matched followed persons (cost); intention-to-follow mortality difference; and as-treated death rate ratio, silence vs follow-up. Performance measures were the bias (estimator − truth, mean ± MCSE over seeds); the coverage, i.e. the share of seeds in which the estimator's within-seed 95% interval contains the truth of the same seed (these intervals use paired standard errors and ignore the reuse of controls, so they are anti-conservative); and failure rates (sign opposite to the truth; relative bias above 50%). The mechanisms are not recalibrated to keep the observed association between rupture and events constant, because that would require the calibration of gate G2. Each mechanism therefore has its own truth, and only the relation between estimator and truth is interpreted.

3. Results

Results come from successive versions of the benchmark, and each is labelled with the version that produced it. For this revision every long-term result (§3.2–3.6) except the single-seed cohort emulation of §3.3 (seed 42, n = 20,000) was re-run over 10 seeds, under v0.17 (costs in 2024 euros; worlds REF, SQ, PA and PB) or v0.19 (heart-failure registry ceiling 0.24; worlds PC and PE; world PE at a remote-monitoring cost of 110 € per equipped month, re-run with v0.19.2 or corrected exactly); the structural experiment (§3.7) used v0.18, whose default structure is identical to v0.17. Daily-trunk results that were not re-run keep their original version label. The provenance of every result (version, replication, population and script) is summarised in the Data and code availability section and detailed in Supplementary Table S6.

3.1 Detection, observability and evaluation designs

Short-horizon detection is a secondary, internal experiment. Historical detection campaigns (v0.2–v0.11; Supplementary Table S9) found that claims alone were close to chance for decompensation within 30 days, and that the sensor that helped depended on the organ: daily IoMT signals for heart failure, dense point-of-care biology for acute kidney injury on CKD. In the internal prediction experiment re-run for this revision (v0.18; not a validation), ten seeds gave pooled test sets of 1,998 equipped patients with claims-known heart failure, observed at 15,730 landmarks, with 176 distinct hospitalisations for acute heart failure (12–24 per seed). The ROC-AUC was 0.57 [95% CI 0.52–0.62] with claims alone and 0.70 [0.65–0.74] with IoMT features (range across seeds 0.37–0.69 and 0.61–0.79). With an alert quota of 5% of patients at each index date, the IoMT model caught 24% of episodes with a positive predictive value of 5.3% per alert (patient-date), about 19 alerts per true episode; repeated alerts for the same patient at successive dates were counted separately. Raw scores were poorly calibrated (slope 0.37), and isotonic recalibration on calibration sets of about ten events per seed did not correct this (slope 0.12; Supplementary Figure S2). These are properties of an injected prodrome detected by one learner on simulated patients, not evidence that a real registry would learn a transferable signature. Alert burden is a design constraint at least as binding as discrimination; in PE, the clinical capacity behind alerts is examined in §3.6.

Historical daily-trunk results. Earlier campaigns (v0.1–v0.14, not re-run) on the ablation of device-wearing counts, informative missingness, equipment coupled to future disengagement and claims-based targeting are reported in Supplementary Table S9-D. The targeting results there are agreement with the simulator's own rupture definition, by a rule built in the same universe, and do not predict performance on the real SNDS.

Participant-based contrasts are biased exactly to the extent of the selection one posits. With ten seeds (v0.18, daily trunk; Supplementary Table S8), the true number of decompensations avoided after re-engagement averaged 9.5 per seed (range 6–18). The individual RCT was unbiased on average (ratio to truth 1.03 ± 0.45) but so imprecise that its estimate had the wrong sign in 3 of 10 seeds. The stepped-wedge overestimated about twofold (2.19 ± 0.81), because exposure is confounded with calendar time, and pre/post comparisons among returners failed (−0.23 ± 0.17). The naive returned-versus-not-returned contrast and its IPTW adjustment were unbiased on average in this experiment (1.02 ± 0.15 and 1.01 ± 0.16; coverage 90%), because return after contact is drawn independently of risk in the daily trunk; the overestimation reported in an earlier version (×1.66–1.71 on three seeds, v0.5–v0.6) is not reproduced and is withdrawn. Together with the long-term result, where the naive attendee-versus-non-attendee contrast overstated the effect about sixfold because attendance was posited to depend on latent burden (§3.5), this shows that the bias of participant-based comparisons is carried by the selection mechanism one assumes, which a real study cannot observe.

3.2 Long-term trajectories, costs and the bias of observational designs

Results in this section come from the monthly trunk (v0.17, 10 seeds × 60,000 persons, about 18,100 meeting the silence criterion per seed) and are reported as mean ± MCSE [min; max] unless stated otherwise. Contrasts are paired simulated contrasts under the assumed model (E1) or model-defined subgroups (E2), except the claims estimators (E3).

A cohort defined by reappearance sees less than half of disengagement. The silence criterion captured 89.4% of true first ruptures. Within 10 years, 51% of the disengaged reappeared through a hospital stay, 24% through an emergency visit only, 19% returned to primary care and 5.7% died without reappearing. Overall, only 46% of true ruptures would enter a cohort defined by hospital reappearance (Figure 2).

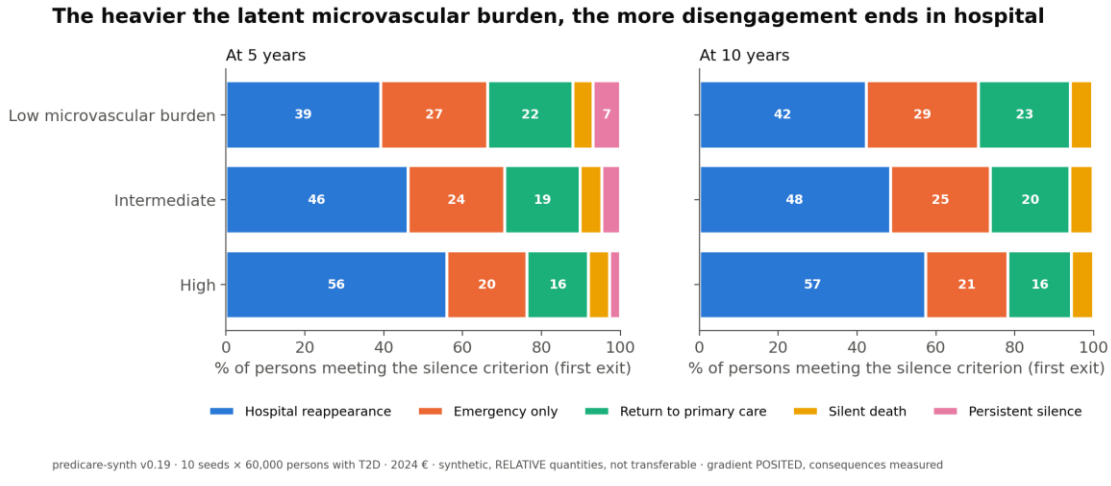

Figure 2 — Competing exits from disengagement by latent microvascular stratum. The gradient on event hazards is posited; its consequences on exits and costs are measured.

The microvascular hierarchy shapes outcomes, but claims do not see it. Hospital reappearance at 10 years rose from 42% (low burden) to 48% (intermediate) and 57% (high). Mortality at 10 years was 33.4% vs 27.4% (low), 38.9% vs 33.0% (intermediate) and 46.1% vs 40.4% (high) in the status quo versus continuous follow-up. Yet only 14.0% of latently high-burden patients were classified as high from claims traces, and low and intermediate strata were indistinguishable (86–87% classified low): non-proliferative retinopathy, albuminuria and grade-1 foot lesions leave no trace in the simulated claims head, by design (in the real SNDS, fundus examinations and albuminuria tests are reimbursed acts, but without their results).

Under the assumed model, disengagement costs lives but not money (Table 5, Figure 3).

Table 5 - Paired simulated differences, status quo minus continuous follow-up, among persons meeting the silence criterion (10-year window, payer perspective, undiscounted; mean ± MCSE [min; max] over 10 seeds). Status: replicated result (10 seeds).

10-year window, SQ − REF (E1)	All strata	High latent burden
Total cost difference	−1,981 ± 76 € [−2,282; −1,480]	−241 ± 160 € [−933; +782]
of which ambulatory care and pharmacy	−9,004 ± 26 €	−9,698 ± 41 €
of which hospital stays	+3,817 ± 14 €	+4,145 ± 15 €
of which dialysis	+3,141 ± 83 €	+5,256 ± 148 €
Mortality (points)	+5.8 ± 0.1 (41.2% vs 35.3%)	+5.7 ± 0.1
Life-years lost in the window	0.42	0.44

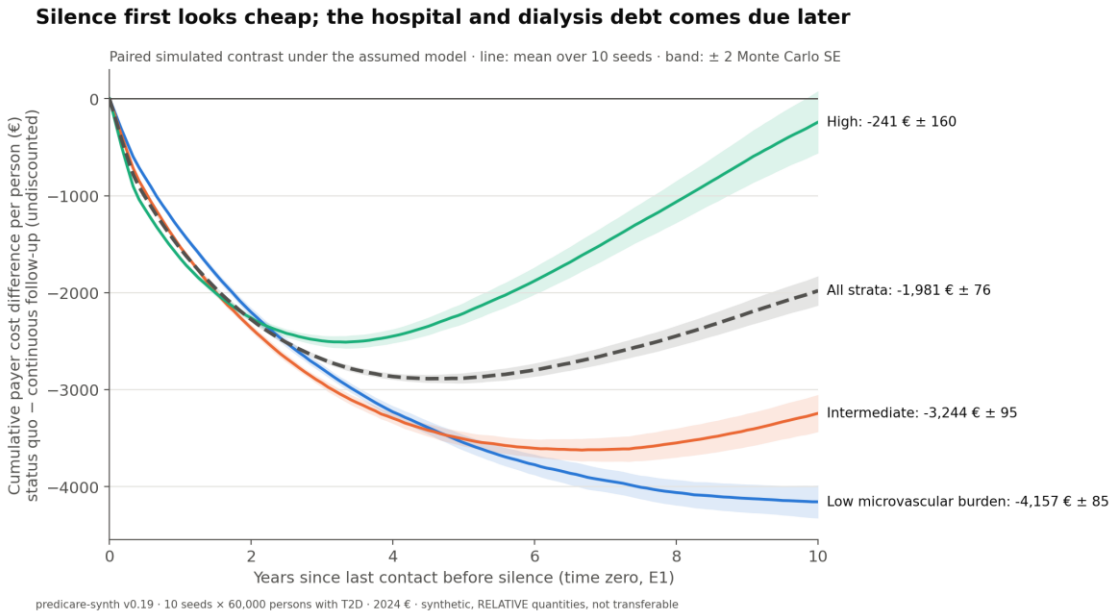

Figure 3 - Cumulative cost difference, status quo minus continuous follow-up, under the assumed model. Silence first looks cheap (care not consumed); the hospital and dialysis debt comes due later. In the high-burden stratum the net 10-year difference is not distinguishable from zero (−241 ± 160 €), but its dialysis and hospital components are robust (+5,256 ± 148 € and +4,145 ± 15 €). Band: ± 2 MCSE over 10 seeds.

Among patients still silent and alive at 5 years, the latent debt was measurable: CKD stage 4–5 in 8.4% vs 6.0%, high stratum 66.8% vs 57.5%, and costs in years 5–10 exceeded continuous follow-up by 1,610 € after a saving of 4,600 € in years 0–5.

Restricting the cohort to reappeared patients overstates the cost of disengagement (Table 6, Figure 4).

Table 6 — Excess cost attributed to disengagement: paired truth versus two claims-based cohort designs (undiscounted; mean ± MCSE over 10 seeds). Status: replicated result (10 seeds).

Excess cost attributed to disengagement	5 years	10 years
Truth (paired simulated contrast, E1)	−2,882 ± 26 €	−1,981 ± 76 €
Claims study: all silent patients vs matched followed	−5,020 ± 85 €	−4,628 ± 212 €
Claims study: reappeared patients only vs matched followed	−1,165 ± 164 €	**+1,106 ± 372 €** [−1,096; +3,442]
Bias of the reappeared-only design (study − truth)	+1,717 ± 154 €	+3,087 ± 323 € [+1,183; +4,922]

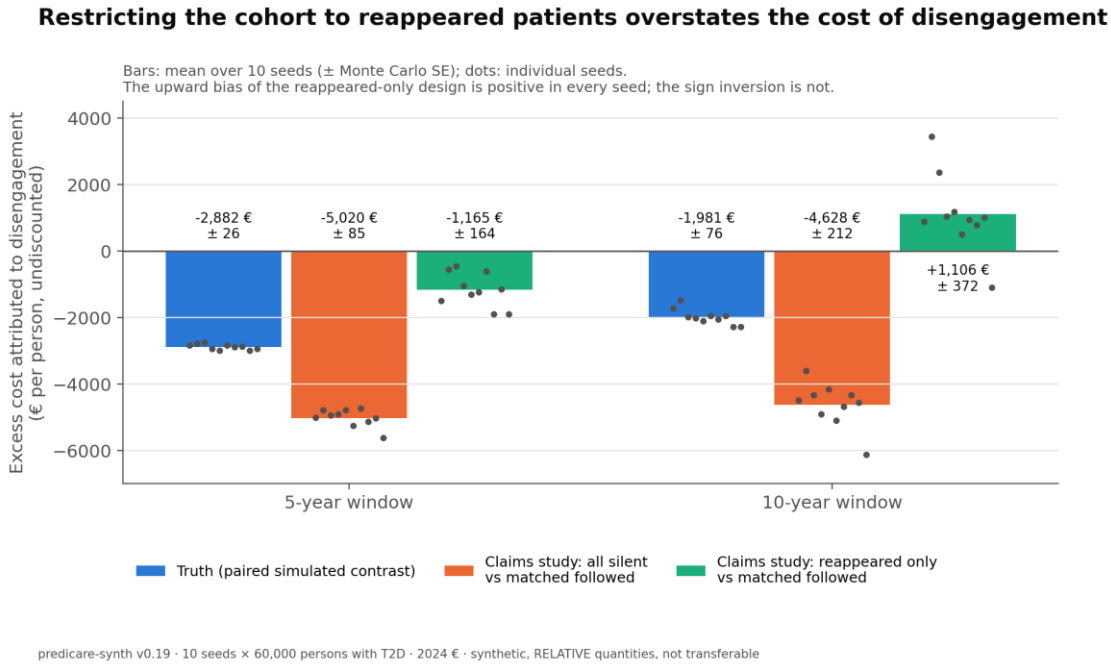

Figure 4 — Selecting on decompensation manufactures part of the excess cost one seeks to show. Bars: mean ± MCSE over 10 seeds; dots: individual seeds.

The upward bias of the reappeared-only design was positive in every seed; the sign inversion itself was not (one seed of ten gave a negative 10-year estimate). The robust finding is the direction and size of the bias, not the inversion.

Sensitivity. Across 13 scenarios (10 seeds × 60,000 each), three results held: excess 10-year mortality (+5.3 to +9.6 points); the cost gap between the high- and low-burden strata, which stayed positive (+3,300 to +5,400 €) and persisted when the posited microvascular gradient was removed (+5,318 €), showing that it is driven by the renal mechanism (CKD ≥ 3b, accelerated progression, dialysis) rather than by the injected gradient; and the upward bias of the reappeared-only design, present in every scenario (+930 € with an

excess decompensation risk ×5 to +6,660 € with ×2). The *sign* of the net cost of disengagement did not hold: it turned positive when the ambulatory fraction of follow-up cost was 30% (+402 ± 75 €), was indistinguishable from zero when episode costs were multiplied by 1.5 (−72 ± 76 €), and approached zero when the excess decompensation risk during rupture was ×5 (−722 ± 79 €). Excess mortality was not monotonic in the decompensation multiplier (6.1 points at ×2, 5.7 at ×5): more decompensations mean more hospital reconnections and therefore shorter ruptures, a property of the posited reconnection mechanism.

3.3 Survival analyses and cost-utility

Cohort emulation (v0.19 export, n = 20,000, seed 42). The claims-only script identified 6,333 disengaged patients and matched all of them at level 1. All SMDs were ≤ 0.005 in absolute value except prior 12-month cost (−0.076), whose residual imbalance remains within tertiles. Ground-truth mortality at 10 years for these patients was 42.2% in the status quo and 36.9% had they stayed followed (+5.3 points).

Each survival analysis sees a different part of the effect (Table 7).

Table 7 - What each survival analysis sees of the mortality effect of disengagement (seed 42, n = 20,000; survival outcomes are identical in v0.14 and v0.19, since the causal model is unchanged). Status: single-seed analysis, not replicated; its 10-seed counterpart for intention to follow is given in §3.7.

Analysis (10 years unless stated)	Disengaged	Matched followed	What it shows
Overall survival, intention to follow (Kaplan–Meier)	60.4%	62.5%	gap of 2.1 points vs 5.3 in truth: contamination — 41% of the followed disengage later, most disengaged return
Free of hospitalisation at 5 years (Kaplan–Meier)	30.0%	45.8%	the gap opens in years 1–3, then curves run parallel
Cumulative incidence of hospitalisation (Aalen–Johansen)	79.0%	69.4%	1 − KM overestimated incidence by ~8 points (86.8% / 77.0%); the between-cohort gap is preserved
Overall survival, as treated (censoring at change of state), 5 years	77.7%	84.1%	mortality gap reappears, under informative censoring
Rate ratio silence / follow-up, decompensation	×2.17	×2.74	silence doubles to triples the decompensation rate while it lasts
Rate ratio silence / follow-up, death	×1.01	×1.15	instantaneous death rate barely changes

The last two rows reconcile the others: mortality from disengagement is **delayed**. Patients decompensate during silence, reappear at hospital, are reconnected and die later, counted as "followed". Any analysis that attributes death to the state observed at death will miss most of the effect; only the matched counterfactual measures it.

Cost-utility (monthly trunk, 20 years, 2.5% discounting; Table 8, Figure 5).

Table 8 - Cost-utility of continuous follow-up and re-engagement versus status quo (monthly trunk, 20 years, 2.5% discounting, payer perspective; mean ± MCSE over 10 seeds). Status: replicated result (10 seeds).

Strategy vs status quo	Δ cost per person	Δ QALY per 1,000	ICER without fixed cost	Headroom at 30 / 50 / 100 k€ per QALY
Continuous follow-up (theoretical bound)	+689 ± 28 €	+175.5 ± 1.8	~3,900 €/QALY	—
Re-engagement, queue A (longest silence)	+89 ± 10 €	+21.3 ± 0.4	~4,200 €/QALY	49 / **88** / 184 €
Re-engagement, queue B (microvascular burden)	+131 ± 11 €	+19.9 ± 0.4	~6,600 €/QALY	42 / 78 / 168 €

Headrooms are expressed per covered person with type 2 diabetes per year (E4); the MCSE of the 50 k€/QALY headroom of queue A was 2 €. At a 10-year horizon this headroom fell to 24 ± 1 € [19; 31], because QALY gains accrue late while operating costs are paid from day one. Across the 13 sensitivity scenarios it

ranged from 43 € (return probability 0.25, or a monthly budget of 0.1%) to 216 € (budget 0.5% per month). Re-engagement was dominant on average (cheaper and more effective; ΔC −25 €) when the ambulatory fraction of follow-up cost was 30%. Adding treatments (v0.13 → v0.14) raised the ICER without fixed cost from ~2,800 to ~4,900 €/QALY, because re-engaged patients resume costly drugs, but left the headroom almost unchanged (93 → 90 €).

The two queues gained the same life-years at population level but not for the same patients: queue B gave lower 10-year mortality in the high-burden stratum (44.9% vs 45.3% for A and 46.1% in status quo; difference −0.39 ± 0.04 points), queue A in the low-burden one (32.1% vs 32.9%; +0.82 ± 0.05 points). With long silences, queue B gained nearly twice as many life-years per disengaged patient (0.085 vs 0.047 years in v0.13).

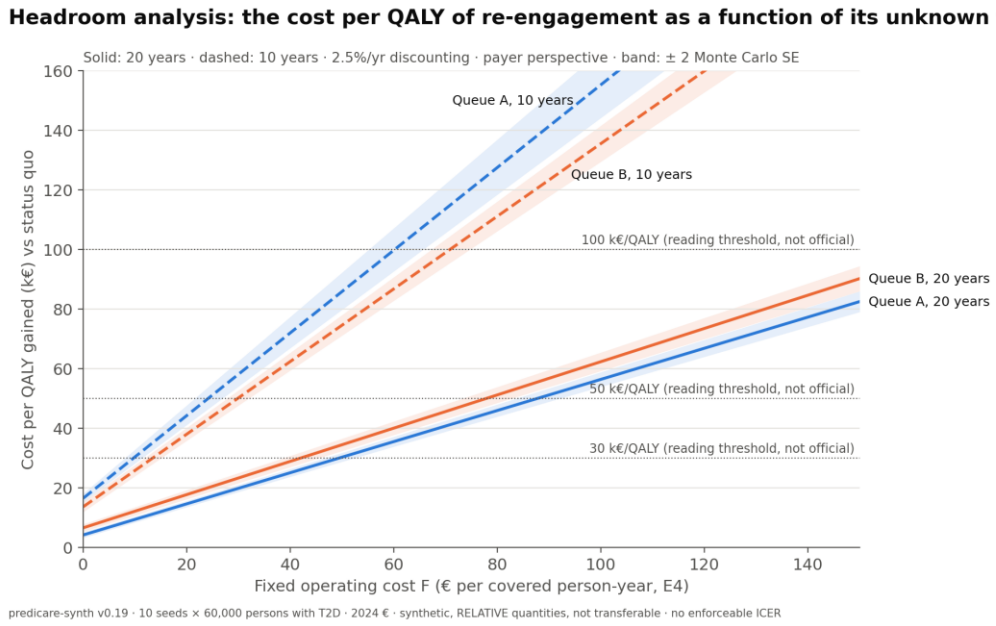

Figure 5 - Headroom analysis: ICER of re-engagement as a function of the unknown fixed operating cost (€ per covered person-year), 10- and 20-year horizons; band ± 2 MCSE over 10 seeds. Thresholds are reading aids, not official French values.

3.4 Crossed destinies, the IoMT variant and probabilistic sensitivity

IoMT variant after re-engagement (PC, v0.19). Equipping re-engaged patients added +0.33 ± 0.14 QALY per 1,000 to re-engagement alone (+21.3 ± 0.4). Its value came mostly from handling the loss of an equipped patient's signal outside the re-engagement budget (+1.57 ± 0.16 with a dedicated telemonitoring queue), that is, from continuity rather than from avoided decompensations; the registry curve of this world rose to a recall of 0.22 after twenty years, against a ceiling of 0.24 (Supplementary Table S11, Supplementary Figure S3). The maximal IoMT fee of PC (9 € per equipped person-month, 32 € with a dedicated telemonitoring queue) does not depend on the kit cost assumed, and it lies below the cost of a continuous glucose sensor alone (about 86 € per month): with a kit that includes one, equipping re-engaged persons is not cost-effective at 50 k€/QALY on average in any scenario of the model.

Two questions, two populations. "How much does disengagement cost?" is asked by the payer when the patient disappears, and the relevant population is all silent patients. "How much do patients who come back damaged cost?" concerns the subgroup that reappears at hospital; the right comparator is the same patients had they stayed followed, not an ordinary followed group. The benchmark computes both (10-

year window, SQ − REF, undiscounted; Table 9). Rows defined by an outcome observed in SQ are model-defined principal strata (E2); only rows defined by the claims-observable stratum describe subgroups that a real study could form.

Table 9 - Cost of disengagement by population, payer perspective (10 years, SQ − REF, undiscounted; mean ± MCSE over 10 seeds; 60,000 persons per seed; collective perspectives in Supplementary Table S12). Status: replicated result (10 seeds).

Population	Share of silent patients	Cost, payer	10-year death, SQ vs REF
All silent patients	100%	−1,981 ± 76 €	41.2% vs 35.3%
Reappeared at hospital	51%	+835 ± 134 € [350; 1,683]	40.7% vs 35.2%
Reappeared at hospital, high latent burden	28%	+2,796 ± 235 € [2,114; 4,295]	45.3% vs 39.9%
High latent burden, all outcomes	49%	−241 ± 160 € [−933; 782]	46.1% vs 40.4%

The usual design (reappeared patients versus ordinary followed patients) reports +1,106 € (Table 6): it overstates the cost of disengagement for this subgroup by about a third (+1,106 € vs +835 €), with wide uncertainty, mixing the price of disengaging with the price of having fared badly. Subgroups defined by an outcome in the status-quo world are principal strata [32]: their comparison with REF is exact under the model and under the chosen coupling of random draws, but not identifiable in real data. In practice, the future of a silent patient is unknown when silence is detected; what can be observed is the microvascular stratum read from claims, under which the probability of hospital reappearance rose from 45% to 61% and 69% and life-years lost from 0.39 to 0.46 and 0.50, while no observable stratum showed an excess cost (−2,280 € to −1,320 €), and only 14% of truly high-burden patients were identified.

Collective perspective. Adding dependency at home or in institutions and informal care (the HAS reference case [25]) deepened the apparent saving for all silent patients, multiplied the excess cost of those who come back damaged, and lowered the headrooms by about 5–13%; the sign of the cost of disengagement was negative in all three perspectives (Supplementary Table S12).

Probabilistic sensitivity analysis. Across 300 joint draws of 23 declared parameter distributions (worlds REF, SQ, PA and PC; 20,000 persons per draw), disengagement was a saving for the payer over 10 years in 95% of draws (per person with diabetes, i.e. averaged over the whole simulated population, of which about a third is ever disengaged: median −613 € [−1,388; +112], discounted) while increasing 10-year mortality in every draw (population level: median +1.6 points [0.9; 2.5]). The headroom of re-engagement at 50 k€/QALY had a median of 83 € [33; 193] over 20 years and 23 € [2; 70] over 10 years; Monte Carlo noise accounted for 7% of its variance. At a fixed cost of 30 € per person per year, re-engagement was below 50 k€/QALY in 98% of draws; at 50 €, 86%; at 100 €, 29%. The return probability after contact dominated (Spearman ρ = 0.65), followed by reachability (0.36), excess mortality during rupture (0.31) and renal acceleration (0.26). The IoMT add-on (base design, without dedicated queue) was positive in 59% of draws only (median +0.28 QALY per 1,000 [−2.1; +2.5]). These distributions are declared plausibility ranges, not uncertainty estimated from data (Supplementary Figures S4 and S5).

3.5 The PREDICARE sequence (world PE)

Over 20 years (v0.19, 10 seeds × 60,000), the invitation budget produced on average 18,840 invitations to 5,760 people, 3,680 people attending at least once, 3,170 people equipped at least once and 1,690 resumptions of conventional follow-up; the prescribed sensitivity for heart failure rose from 0.06 at one year to 0.18 at five years, faster than in PC because equipped disengaged patients decompensate more often and therefore feed the curve faster. The scenario in which monitoring has no effect on the excess risk of disengagement (attenuation 0) is the reference: the attenuation is unmeasured, and any positive value is an assumption that no current evidence supports. The counts above are for this reference (§3.6 uses an attenuation of one half, hence 3,110 equipped); a person may attend at several disengagement episodes, so the ever-equipped share of attendees (86%) exceeds the per-attendance acceptance of 0.80. Headrooms by scenario are given in Table 10 and Figure 6 (left).

Table 10 - PREDICARE sequence (PE − SQ) by scenario: QALY gain and headroom (20 years, discounted; mean ± MCSE over 10 seeds; remote monitoring at 110 € per equipped month, other costs in Supplementary Table S14). Headrooms are a conditional illustration under posited, uncalibrated parameters, not cost estimates or budgets. Status: replicated result (10 seeds), conditional on posited programme parameters.

PE scenario (20 years, discounted)	Δ QALY per 1,000, PE − SQ	Headroom at 50 k€/QALY (€/covered person-year)
Reference: monitoring without effect on excess risk (attenuation 0)	+7.0 ± 0.3	13 ± 1
Attenuation 0.5	+9.4 ± 0.4	28 ± 2
Attenuation 0.5, no selection on attendance	+10.8 ± 0.3	33 ± 2
Attenuation 0.5, sensor attrition 1% per month	+11.3 ± 0.4	34 ± 2
Attenuation 1 + treatments renewed by telemonitoring	+12.5 ± 0.5	47 ± 2
Same, attrition 1% per month	+17.1 ± 0.5	70 ± 2
Comparator: re-engagement alone (PA, 50% return after contact)	+21.3 ± 0.4	88 ± 2

The value of the PREDICARE sequence depends on two quantities nobody has measured — how much remote monitoring offsets the excess risk of disengagement, and whether treatment continues outside conventional follow-up. PE and PA are not compared at equal capacity: PA posits that a contact re-engages 50% of those reached, whereas PE chains reachability, attendance and a 35% resumption after assessment, with different budgets, access rates and follow-up intensities. The comparison is an illustration of scenarios, not a ranking of strategies; harmonising capacities, and measuring these rates, is precisely what a demonstrator must do.

Probabilistic sensitivity of the PREDICARE sequence. Because the scenarios above are chosen, the probabilistic analysis was repeated on worlds REF, SQ, PA and PE (300 draws, 20,000 patients each), adding seven distributions for the sequence to the 23 previous ones: attendance if reached (beta, mean 0.50), resumption of follow-up after assessment (beta, mean 0.35), sensor attrition while disengaged (lognormal, median 4% per month), reconnection after a true alert (beta, mean 0.50), attenuation of the excess risk by remote monitoring (uniform 0–1), renewal of treatments by the telemonitoring team (Bernoulli 0.5) and cost of the T0 assessment (uniform 120–250 €); the remote-monitoring cost is drawn uniformly

between 70 and 150 € (each published draw mapped at the same quantile, costs corrected exactly; Supplementary Table S14). Over 20 years, PE gained a median of +8.8 QALY per 1,000 [2.9; 19.6], and its headroom at 50 k€/QALY had a median of 25 € per covered person-year [−1; 78] (10 years: 1 € [−12; 26]); Monte Carlo noise accounted for about 12% of its variance (within-draw variances computed under the previous cost distribution, not re-estimated at 110 €). At a fixed cost of 10, 20, 30 and 50 €, PE remained below 50 k€/QALY in 84%, 64%, 39% and 15% of draws. The headroom depended chiefly on the attenuation of excess risk (Spearman ρ = 0.58), attendance (0.39) and resumption of follow-up (0.32), then on excess mortality during rupture (0.27), renal acceleration during rupture (0.20) and spontaneous return (−0.20). These results quantify the dependence of PE on unmeasured quantities; they do not make its value better established than the scenarios suggest, and they identify what must be measured rather than a target budget. Units: the headroom is expressed per person with type 2 diabetes in the territory per year (E4; discounted person-years of the whole simulated population, about 11 per person over 20 years), not per targeted, invited or equipped patient. Translations of the same total per invited person, attendee or equipped patient-year are given in Supplementary Table S10; they redistribute a conditional quantity over a smaller denominator and are not budgets.

Value of information. At 50 k€/QALY and a fixed cost of 30 € per covered person-year, where the sequence was cost-effective in 39% of draws, the EVPI was 86 ± 7 € per person with diabetes over 20 years. The EVPPI was 52 ± 8 € for the attenuation of excess risk under monitoring, 33 ± 8 € for attendance, 26 ± 7 € for excess mortality during rupture, 18 ± 6 € for resumption of follow-up, and 63 ± 7 € for attenuation, attendance and resumption jointly; reachability (10 ± 6 €), renal acceleration and spontaneous return during rupture (14 ± 7 € and 14 ± 6 €) and sensor attrition (11–12 ± 6–7 €) lie at about two standard errors and are not interpreted, nor is any other parameter. At 10 € the decision was no longer nearly certain (cost-effective in 84% of draws; EVPI 13 ± 2 €). With the previous remote-monitoring cost distribution (10–40 €), the EVPI was 41 € and the EVPPI of attenuation 13 € (Supplementary Table S14): a realistic kit brings the decision closer to indifference and doubles the value of information. The ranking supports collecting evidence first on attenuation (through intermediate outcomes under monitoring, with the identifying assumptions this requires) and on attendance; the values are per person over 20 years, and a population EVPI would require the size and duration of the decision, which we do not posit.

Selection, and which population the offer is evaluated in. Results are for 10-year mortality in the reference scenario without attenuation (mean ± MCSE over 10 seeds; Supplementary Table S10). Among all eligible persons (E5; about 18,100 per seed, the population of E1), the offer reduced 10-year mortality by 0.31 ± 0.02 points and raised the 10-year payer cost by 764 ± 14 € per eligible person (337 € at the code default of 25 € per equipped month). Under the monthly invitation budget, only 28.0% of eligible persons were ever invited, after a median of 41 months, and 18.3% attended; eligible persons never invited had, by construction, identical outcomes in both worlds. The complier average causal effect [18] (ITT divided by attendance, since no one attends in the status quo) recovered the truth among attendees, −1.72 ± 0.09 points. Among the persons actually invited (time zero: first invitation; a population selected by the queue and the budget), the effect of the offer was −1.25 ± 0.09 points, and 67.1% attended after repeated invitations. The naive comparison of attendees with non-attendees among the invited gave −11.1 ± 0.5 points, about six times the truth among attendees with the same time zero (−1.87 ± 0.13 points) and nine times the effect among the invited (Figure 6, right), because non-attendees were sicker (63% vs 47% high latent burden; 46% vs 37% status-quo 10-year mortality). With an attenuation of one half, the effect of the offer was −0.41 ± 0.02 points among all eligible persons and −2.28 ± 0.12 points among attendees. The CACE recovers the truth here only because, by construction, the offer has no effect on non-attendees (exclusion restriction) — an assumption to be argued in the field, where an invitation may change care use even without attendance (a reminder effect); the demonstrator should therefore measure care use among non-attendees. The effect of the offer among all eligible persons is therefore the primary estimand: in the

simulation it corresponds to design (i) below, an evaluation randomising territories before targeting starts with the same eligibility rule in both arms and capacity fixed locally. Repeated invitations are part of the policy evaluated. Because the budget dilutes the offer, 10-year mortality is not a feasible primary endpoint. Among about 18,100 eligible persons per 60,000 people with type 2 diabetes, a status-quo 10-year mortality of 42.5% corresponds to about 7,700 deaths, of which the simulated policy avoids about 57. Detecting a difference of 0.31 points with 80% power at a two-sided 5% level would require about 400,000 eligible persons per arm under individual randomisation (normal approximation for two proportions), before any inflation by the design effect of a cluster trial; restricting the analysis to the persons actually invited (−1.25 points) would require about 24,000 per arm, but in a population selected by the queue. These orders of magnitude are conditional on posited, uncalibrated parameters. They indicate that a demonstrator would need intermediate endpoints (resumption of follow-up, treatment continuity, decompensations), whose effect sizes the benchmark posits rather than estimates, and should report the invited and attending fractions alongside the ITT. The recommendation is a direction for a protocol, not a protocol: the unit of randomisation, the number of clusters, the endpoints and the sample size remain to be designed and justified with real data. The effect among participants requires the exclusion restriction (monotonicity holds by design here, since attendance is impossible without invitation) and is secondary. The demonstrator should pre-specify the unit of randomisation, the eligibility rule and time zero, the handling of repeated invitations, and, for clustered designs, the clusters, the periods, contamination between arms and secular trends [33].

What a trial of the offer would estimate under a shared capacity. The offer is delivered through a monthly queue with a fixed budget, so an invitation sent to one person is a place not sent to another: assignments and outcomes are not independent across persons (interference). Three designs are possible, and they do not estimate the same effect. (i) *Randomisation by territory* (or by primary-care cluster), allocated before targeting starts, with the same eligibility rule and time zero in both arms and each arm's capacity fixed locally: this estimates the effect of the policy as a whole, budget, queue order and selection included, among all eligible persons. This is what E5 represents in the simulation, and the −0.31 point is the effect of this simulated policy under its budget; it is not the effect of an invitation received by each eligible person. (ii) *Individual randomisation of invitation places* among the persons eligible in each period (a lottery within the queue): this estimates the effect of being offered a place, among those competing for places in that period, conditional on the capacity; persons not drawn may be drawn later, so the handling of re-entry into the lottery and of repeated invitations must be pre-specified, and the contrast is diluted as controls are later invited. (iii) *Comparison of allocation rules under the same budget* (for example the two queue orders of §3.3, randomised by cluster): this estimates the difference between rules, not the effect of the programme. Whatever the design, the protocol must fix before assignment who enters the denominator, at which time zero, how re-invitations are counted, and how capacity is shared between arms.

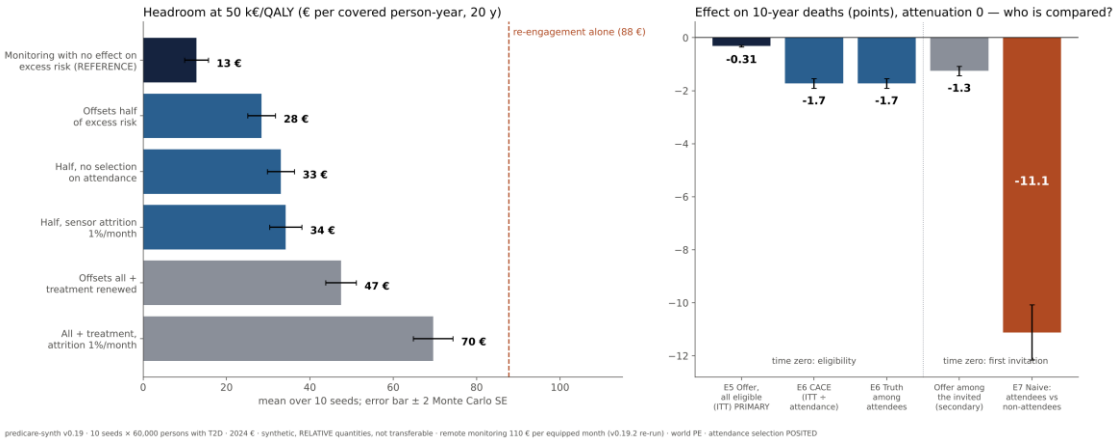

Figure 6 - World PE. Left: headroom at 50 k€/QALY under the two unmeasured quantities, remote monitoring at 110 € per equipped month (bars: mean ± 2 MCSE over 10 seeds; dark bar: reference, attenuation 0). Right, reference scenario: the effect of the offer among all eligible persons (E5, primary), its scaling by attendance (CACE, E6) and the truth among attendees (E6), all with time zero at eligibility; then, with time zero at first invitation, the effect among the invited and the naive attendee-versus-non-attendee contrast (E7).

3.6 A T0 assessment that decides, and the clinical capacity behind alerts (v0.16 mechanisms, v0.19 runs)

Two mechanisms were added to PE (attenuation of excess risk 0.5, so that monitoring has an effect for alerts to act on; 10 seeds × 60,000). First, the T0 assessment now measures CKD, heart failure and a high microvascular stratum with posited errors and can drive equipment: under a stratified policy, only attendees classified at high risk are equipped. Second, alerts are handled by a team of limited capacity (alerts reviewed per 100 equipped patients per month, against an alert budget of 5), possibly prioritised by T0 risk, and avoidability decreases linearly with the response delay up to the 13-day median lead time of the daily trunk (Table 11).

Table 11 - T0 equipment policy, assessment errors, alert capacity and response delay in world PE (attenuation 0.5; remote monitoring at 110 € per equipped month; mean ± MCSE over 10 seeds; equipped persons and avoided decompensations are totals per 60,000 persons over 20 years). Status: replicated result (10 seeds), conditional on posited programme parameters.

Scenario (PE − SQ, 20 years)	Δ QALY per 1,000	Headroom at 50 k€/QALY	Equipped patients	Decompensations avoided by alerts
Reference (all attendees equipped, unlimited capacity, 2-day delay)	+9.4 ± 0.4	28 ± 2 €	3,110	97
T0-stratified equipment, perfect assessment	+8.4 ± 0.3	29 ± 1 €	2,085	69
T0-stratified, assessment errors (sensitivity 0.80 / 0.70 / 0.60, 5% false positives)	+8.1 ± 0.3	28 ± 1 €	1,861	62
Capacity 3 alerts per 100 equipped per month	+9.2 ± 0.4	28 ± 2 €	3,106	59

Capacity 1.5	+9.3 0.4	±	28 ± 1 €	3,111	28
Capacity 1.5, alerts prioritised by T0 risk	+9.3 0.4	±	28 ± 2 €	3,108	35
Response delay 7 days	+9.4 0.4	±	28 ± 2 €	3,111	57
Delay 7 days and capacity 1.5	+9.3 0.4	±	28 ± 1 €	3,108	16

Capacity and delay removed up to 84% of alert-driven avoidances (97 → 16), and prioritising by T0 risk recovered about a fifth of the avoidances lost when capacity fell from 3 to 1.5 alerts per 100 equipped patients per month (59 → 28 → 35), i.e. 10% of the loss relative to unlimited capacity, yet QALYs and the headroom barely moved: all differences were within about two MCSE. In this model, the value of the sequence therefore does not run through alerts — they avoid about a hundred decompensations among 60,000 people over 20 years — but through the posited attenuation of excess risk during monitored disengagement and through reconnection to care. This is partly structural, since attenuation applies to every monitored patient whatever the alerts, and it makes that parameter, not the detector, the quantity on which a demonstrator should collect intermediate evidence first; a randomised offer would not, by itself, attribute its effect to attenuation. A T0-stratified equipment policy equipped a third fewer persons (2,085 vs 3,110) but lost about a tenth of the QALY gain (8.4 vs 9.4 per 1,000), because lower-risk attendees also benefited from monitoring under the posited attenuation; assessment errors cost little more. The headroom of stratification depends on the equipment cost: at the code default of 25 € per month it was about 4 € lower than equipping all attendees (36 vs 40 €); at 110 € the two policies are equivalent (29 vs 28 €), and at 150 € stratification is preferable (25 vs 23 €), because equipping lower-risk attendees then costs more than the QALYs it gains. The comparison would also shift if monitoring of low-risk patients brought no attenuation.

Sensitivity to the registry ceiling. Halving or quartering the heart-failure ceiling (0.12 and 0.06 instead of 0.24; 10 seeds each) left the PE headroom unchanged to within 0.2 € (24.6 → 24.5 € without attenuation; 40.1 → 40.0 € with an attenuation of one half; values at the code default of 25 € per equipped month, not re-run at 110 €) and the QALY gain unchanged to the first decimal, while decompensations avoided by alerts fell by 10–15% (128 → 109 without attenuation; 97 → 83 with); in PC, the bound on the IoMT fee moved from 9 to 8 € per equipped person-month (Supplementary Table S9-C). Avoidance does not fall in proportion to the ceiling because alerts for acute kidney injury and diabetic decompensation keep their own ceilings and because the recall floor is chance at the quota (0.05). Within this range, the economic results are insensitive to how the daily-trunk recall is transferred to the monthly queue; the number of alert-driven avoidances is not, and it would require simulating the alert chain itself (repeated alerts, deduplication, case mix) to be interpreted.

3.7 The claims designs under alternative data-generating mechanisms

Results come from v0.18 (14 mechanisms × 10 seeds × 60,000 persons; Figure 7, Supplementary Table S7).

Selection on reappearance biased the cost upwards under all 14 mechanisms. The mean bias of the reappeared-only design ranged from about +2,600 € (M4) to +7,500 € (M3) per person over 10 years. It was positive in 139 of the 140 mechanism–seed pairs, and the coverage of its 95% interval was at most 40%. Whether the estimate also inverted the sign depended on the mechanism: it did, on average, when disengagement was causal and patients reconnected after hospital (9 of the 13 mechanisms with a clearly

negative truth; in M4, M7 and M8b the estimate is within about 1.5 MCSE of zero); it did not when disengagement was mainly a marker of latent severity (M1–M3) or when reconnection was absent (M8); and under partial and intermittent disengagement (M6) the truth is close to zero (−70 ± 60 €) while the design reports +4,300 €. The robust finding is the direction of the bias, not the inversion.

Comparing all silent persons with matched followed persons overstated the saving under every mechanism, by about 1,300 € (M8) to 3,000 € (M4) on average, and in 139 of 140 mechanism–seed pairs. Matching on claims covariates removed neither the loss of paid care nor the latent differences.

The direction of the intention-to-follow mortality bias depended on the mechanism. When disengagement was causal (M0, M5–M11), intention to follow diluted the effect, by about 1.7 (M6) to 4.2 (M8) points (under the reference mechanism M0, 3.3 ± 0.1 points against a truth of 5.8 ± 0.1 over 10 seeds; 2.1 against 5.3 in the single-seed cohort emulation of Table 7), through contamination and return to care, as in §3.3; when a causal effect was combined with confounding by a shared frailty (M4), it overstated the effect by 1.5 points. When disengagement was partly or wholly a marker of latent severity, the same design overstated it: 8.7 points against 3.4 under M1, and 6.5 points of excess 10-year mortality under M2 and M3, where the true effect was 0.1 point. A claims study therefore cannot tell, from its own data, whether its mortality contrast is too small or too large.

The as-treated death rate ratio never recovered the 10-year effect. It ranged from 1.06 (M10) to 1.57 (M8) under causal mechanisms and was 0.98–0.99 under non-causal ones: it measures the concurrent risk during silence, not the delayed mortality that follows it.

The economic argument is conditional on the causal fraction. The headroom of re-engagement alone (PA, 50 k€/QALY, 20 years) was 88 € per covered person-year under M0; it fell to 28 € when half the excess risk was causal (M1) and to 27 € when 30% of silences were care outside the system (M7); it became negative (down to −52 €) when disengagement was a pure marker (M2, M3), because re-engagement then brings back costly care without health gain; and it rose to 173 € when reconnection after hospital was absent (M8). The quantity that decides whether a vigilance programme can be worth its cost is thus the causal part of the excess risk of disengagement. The benchmark cannot supply it, and neither can a randomised evaluation of the offer on its own. Such a trial identifies the effect of the offer — targeting, invitation, assessment, monitoring and any resumption of care taken together — in the randomised population, whatever the confounding between disengagement and severity; it does not identify the share of the excess risk caused by disengagement itself, nor separate the contributions of treatment resumption, the T0 assessment, monitoring and alert response. Attributing a trial effect to one of these mechanisms requires additional measurements and identifying assumptions (intermediate outcomes such as treatment continuity and decompensations under monitoring, multi-arm or factorial designs, or mediation analyses with their own conditions).

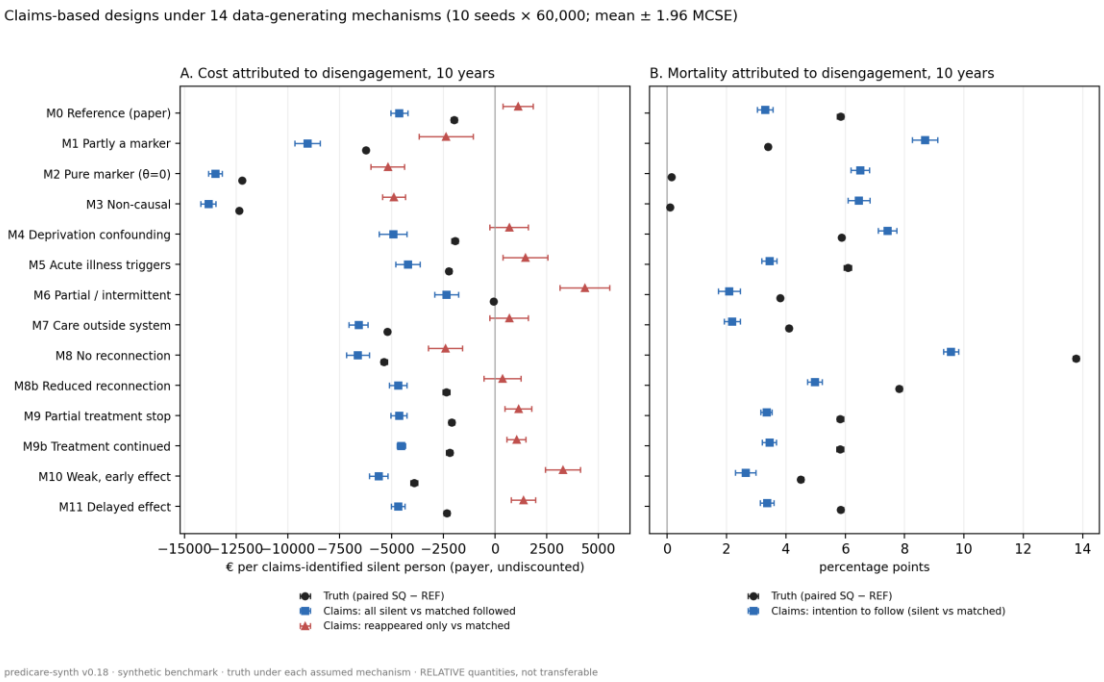

Figure 7 - Claims-based designs under 14 data-generating mechanisms: paired truth (SQ − REF under each mechanism) and claims estimators, mean ± 1.96 MCSE over 10 seeds. A: cost attributed to disengagement at 10 years; B: mortality at 10 years (intention to follow). Each mechanism has its own truth; only the relation between estimator and truth is interpreted.

4. Discussion

Principal findings. In a synthetic benchmark where truth is known by construction, the posited disengagement mechanism produced about 6 points of excess 10-year mortality while being roughly neutral, or saving, for the payer. These magnitudes restate the injected hazards; the findings that do not depend on their exact values are the mechanisms and the design biases that follow. In the model, the saving is not a paradox: silence suppresses useful care that is also paid care, and its harm is paid later in lives and, for the most damaged patients, in dialysis. In the configurations tested, the three claims-based designs tested each failed in a characteristic way: selection on reappearance biased the cost upwards, to the point of inverting its sign on average, intention-to-follow survival diluted the mortality effect through contamination, and as-treated attribution missed a mortality that is delayed beyond the silence itself. In the PREDICARE sequence, the prescribed registry curve (an imposed performance gain, not a simulated learning process) rose quickly because monitored disengaged patients decompensate often, but the clinical value depended on two unmeasured quantities — how much remote monitoring offsets the excess risk of disengagement, and whether treatment continues outside conventional follow-up — and the posited selective attendance was enough to make any participant-based comparison misleading. Limiting the capacity of the team that handles alerts, or lengthening its response, removed most alert-driven avoidances without changing QALYs: in the model, value ran through the attenuation of excess risk, not through the detector.

Properties of the designs, properties of the mechanism. The structural experiment separates two kinds of findings. The upward bias of reappearance-based cohorts and the overstated saving of matched claims comparisons persisted across all 14 mechanisms tested; they are properties of these designs within this class of scenarios. That class is explicit and author-defined: it is neither a representative sample of plausible

mechanisms nor a probability distribution over them, and it does not show that the direction of these biases holds for SNDS studies in general. Mechanisms not tested include a probability of reconnection or return that depends on latent frailty, time-varying confounding between deprivation, observation and outcomes, heterogeneous causes of rupture with different effects, and territorial differences in coding and recording practices. The dilution of mortality by intention-to-follow analysis, the magnitude of every contrast and the economic headroom did not persist; they are properties of the assumed mechanism, and above all of the causal share of the excess risk of disengagement.

Algorithmic findings.

1. *In the simulated world, the information needed for 30-day prediction is absent from claims.* This is a property of the observation architecture, not of the model: the same gradient-boosting learner on the same patients moved from near chance (ROC-AUC 0.57) to 0.70 when daily signals were added in the internal experiment — a discrimination gain on simulated prodromes, not a predictive validation: the pooled test sets held 176 episodes, the positive predictive value at a 5% quota was 5.3%, and calibration was poor.

2. *Historical campaigns (Supplementary Tables S9-A and S9-D).* In the simulated prodromes the useful sensor depended on the organ, and routine biology failed because of its cadence, not because the signal was absent; under informative missingness the predictor migrated to a behavioural proxy while AUC rose; and a social gradient in equipment halved the coverage of disengaged patients without changing AUC among the equipped. AUC does not certify a predictor, nor its equity: a governance check on feature attribution and on coverage is needed, not a better score.

3. *Priority rules are bounded by observability.* A queue ordered by microvascular burden shifted benefit towards high-risk patients but could not outperform a simpler rule overall, because simulated claims, by construction, identify only 14% of high-burden patients.

4. *Participant-based estimators are only as good as the selection one assumes.* When return to care was independent of risk, the naive contrast and its IPTW adjustment were unbiased; when attendance depended on latent burden, the naive contrast overstated the effect about sixfold. Pre/post comparisons failed, a stepped-wedge overestimated about twofold through calendar confounding, and an individual RCT was unbiased but had the wrong sign in 3 of 10 seeds for rare events — which argues for a pre-registered randomised evaluation of the offer, with cluster or stepped-wedge variants only under randomised sequences and explicit modelling of calendar time, and for more frequent intermediate endpoints.

Relation to prior work. Synthetic health data are usually produced either by learning from records (generative networks, language models), which reproduces their observation regime, or by rule-based simulators such as Synthea [34], which target realism of individual records. Our generator belongs to the second family but serves a different end: it is a *test bench for designs*, closer to plasmode simulation [35] and to simulation studies of statistical methods [3], with common random numbers to reduce variance between counterfactual worlds. The direction of our injected effects follows published evidence (microvascular burden and cardiovascular risk; CVOT meta-analyses for GLP-1 RA and SGLT2i; telemonitoring prodromes in heart failure), but the benchmark does not test these effects; it tests what follows from them. The Health Data Hub synthetic "top diabète" release reproduces the SNDS format, not patients, and carries no realistic longitudinal rupture; a deep generative model trained on it would learn another generator's distribution. Learned generators become relevant once real SNDS data are accessible within their legal framework (gate G4): to calibrate and check the mechanistic generator, or in a hybrid design that keeps the causal trunk for ground truth and learns only the claims observation head.

Limitations.

- *Circularity.* The benchmark predicts what it has injected. All results are relative (between arms, designs or worlds at identical truth), never absolute performance, incidence or cost. This is stated in machine-readable validity domains. The automated tests verify internal invariants (separation of truth, determinism, identity before rupture), not the clinical correctness of the mechanisms; differences between worlds are more defensible than their absolute values, but remain dependent on the injected hypotheses. We tested 14 alternative structures rather than a single one (§2.7, §3.7); they remain posited, uncalibrated and not exhaustive (for instance, no mechanism lets the probability of reconnection depend on latent frailty).

- *Parameters.* Almost all parameters are at status *hypothesis*; calibration against registered sources (G2) is pending. The most consequential and least documented are the return probability after contact (re-engagement) and, for the PREDICARE sequence, the attenuation of excess risk under remote monitoring, attendance and resumption of follow-up. The heart-failure ceiling of the registry curve is taken from the internal prediction experiment (0.24); the earlier, optimistic value (0.39) changes no economic quantity beyond Monte Carlo error but raises alert-driven avoidance by about 12%. A first pass on simulated aggregates (Table S2) suggests that several event rates, notably dialysis start and myocardial infarction, are high relative to French orders of magnitude; this inflates absolute values and can amplify differences between worlds.

- *Costs and utilities.* Reimbursed amounts are tariffs, not production costs (HAS guidance prefers French production costs, such as the national hospital cost study, with tariffs as a fallback [25]); the ambulatory fraction of follow-up cost (0.60) is posited and is the most influential cost parameter for the sign of the cost of disengagement; programme unit costs are posited and will be replaced by the demonstrator's costing; several other items are posited; utilities are not French EQ-5D-5L values; the collective perspective rests on posited transitions to dependency and informal-care hours, without feedback on mortality or QALYs; the probabilistic sensitivity analysis uses declared plausibility ranges, not distributions estimated from data, and does not raise the epistemic status of any parameter.

- *Status of the economic results.* The headrooms answer a formal question (which fixed cost would be compatible with a reading threshold under these assumptions), not the practical question of what PREDICARE may cost; they are a secondary illustration, to be recomputed after calibration (gate G2) and a costing of the demonstrator, and should not be read as an operational order of magnitude. The remote-monitoring cost is partly posited: only its glucose-sensor component (about 86 € per month) rests on a published reimbursement base, which currently covers insulin-treated persons only; at 25 € per month the PE headrooms are about 7 to 18 € higher (Supplementary Table S14).

- *Structure.* A single entry condition (type 2 diabetes) in the long-term trunk; no unflagged relocations (the daily trunk includes them; Supplementary Table S9-D); perfect adherence while followed; no claims signal preceding rupture (pre-index trajectories are flat), so the anticipation of disengagement itself cannot be tested. Type 2 diabetes is the first use case: the broader claims phenotype of the metabolic syndrome is not yet formalised, and the results do not extend to cardio-reno-metabolic patients without a diabetes long-term-condition record. In the base case the T0 assessment reveals CKD and heart-failure status without measurement or classification error; v0.16 adds posited measurement errors and lets T0 drive equipment and alert prioritisation, but not yet monitoring intensity or alert thresholds. The clinical capacity behind alerts is now represented by a capacity, a response delay and a prioritisation rule; the heterogeneity of responses (call, teleconsultation, visit, treatment change), alert fatigue and threshold learning by the team are not modelled. In the main

analyses rupture is a single mechanism with common multipliers; the structural experiment varies its causal share, confounding, partial and intermittent forms, care outside the observed system, reconnection and treatment continuation, but not the diversity of its causes (deprivation, refusal, relocation, access barriers, adverse effects, intercurrent acute illness, depression or cognitive decline), symptom-triggered returns, graded adherence or delayed resumption of treatment. This simplification can make the chain rupture → progression → death sharper than it is. Residual confounding in the claims-based designs is, by contrast, a result rather than a flaw: the benchmark shows that the latent microvascular burden is largely invisible in claims, and real data carry further unrecorded determinants (social position, frailty, preferences) that no observed-covariate adjustment can neutralise.

- *IoMT chains and trunk coupling.* Since v0.15, two distinct chains are carried by the long-term worlds: in the IoMT variant PC, re-engagement → equipment → detection → avoidance, with monitoring stopping at the next rupture; in the PREDICARE sequence PE, targeting → invitation → mobile-unit assessment → monitoring during disengagement → alert and response. PC is a variant, not the intended functioning of PREDICARE. In both, the coupling with the daily trunk is parametric (its performances enter as parameters), not dynamic: no daily physiology is simulated over 20 years, and the registry curve is posited: model training, the choice of labels, temporal validation and transfer across territories are not simulated.

- *Reporting gaps.* Undiscounted cost-utility results and a cost-effectiveness plane are not reported; the probabilistic analysis covers neither alert capacity nor response delay, which were explored deterministically only; daily-trunk results other than the prediction experiment and the evaluation designs were not re-run for this revision and keep their original version labels (Supplementary Table S6).

Implications. For evaluation: the bias of a design should be measured on a synthetic counterfactual before the design is imposed on patients, and because cohorts defined by reappearance overstated the cost of disengagement in every mechanism tested, claims-based studies should not rely on them without an explicit argument that the mechanisms at work differ from those tested; a demonstrator should randomise the offer — by clusters (territories or practices) allocated before targeting starts, each with its own capacity, or by a lottery of invitation places within each period — pre-specify which effect the chosen design estimates under the shared capacity (§3.5), keep the denominator of all eligible people and report the effect of the offer separately from the effect among participants, since attendance is likely to be selective. A stepped-wedge rollout is a conditional option only: in the benchmark, a stepped-wedge overestimated avoided events about twofold through calendar confounding, so it would require randomised sequences, enough clusters and periods, explicit modelling of calendar trends and a plan for contamination, none of which is demonstrated here. For economics: a vigilance programme should be argued as a cost per QALY under an explicit headroom, conditional on the causal share of the excess risk of disengagement, not as a saving. For design: organisation matters more than detection in the PC variant (Supplementary Table S11), equipping re-engaged patients paid off mainly when their signal losses were handled outside the re-engagement budget, and the value of IoMT lay more in continuity than in avoiding decompensations; in the PREDICARE sequence PE, value depended first on how much remote monitoring offsets the excess risk and on attendance, while alert capacity and delay barely mattered and the choice between queue orders is a policy decision about who is protected, which belongs to elected and clinical authorities, not to the model.

Technology and intervention. The benchmark separates what the technology provides in the model, visibility and alerting, from what belongs to the complex intervention around it: adherence, continuity of

treatment and reconnection to care. Under the posited mechanism, the headroom of the PREDICARE sequence varied about fivefold (13 € to 70 € per covered person-year) with the attenuation of excess risk, the renewal of treatments and sensor attrition, whereas capacity limits and response delays that removed up to 84% of alert-driven avoidances left it practically unchanged (§3.6). Within this model, the economics of a vigilance programme thus depend far more on what the care team does with and for the patient than on algorithmic performance. This is a property of the simulated structure, not a demonstration: whether remote monitoring can produce such an attenuation at all, and through which human support, cannot be read from a randomised offer alone: the offer's effect bundles every component of the programme, and its attribution to attenuation would require intermediate measurements and explicit assumptions.

Coverage of the registry. In the reference scenario, 18.3% of eligible persons ever attended the mobile unit and 86% of attendees were ever equipped: the registry would observe roughly one eligible person in six. That sixth is selected by reachability, by attendance — posited to fall with latent burden, so that the sickest are under-represented — and by acceptance and wear of the kit; people who decline equipment, possibly those whose disengagement is deepest, are not followed in the same way. The number of measurements a registry collects therefore says nothing about whether it represents the targeted population, nor whether it holds enough well-dated events to learn robust signatures. This one-in-six figure is a simulated cumulative coverage under posited rates of reachability, attendance, equipment and attrition; it says neither how many persons will have a series complete enough to use, nor how many confirmed events each subgroup will contain. What a real registry must report first, before any predictive result, is its usable longitudinal coverage: duration of wear, missing data, dated and confirmed events, and differences in coverage by level of risk and by social and territorial characteristics.

What the benchmark says, and does not say, about the programme's purpose. PREDICARE is meant first to make a disengaged population observable and to build a registry from which deterioration patterns would be learned retrospectively, and only later to escalate patients to acute care (§1). These steps are a research programme still to be carried out, not capabilities tested here; read against that sequence, step by step, the benchmark informs some links and leaves others untouched.

- *Registry constitution* (which targeted patients are reached, equipped and remain measurable): partly modelled, through attendance, equipment uptake, posited sensor attrition and the handling of signal loss (§3.4, §3.5); targeted patients who decline equipment are not followed, whereas a real registry should document them too.

- *Events linked to the series* (dated, graded, clinically confirmed decompensations, and comparable event-free periods): the model knows every event by construction; what limits learning in practice is the number and diversity of well-dated events, not the number of measurements, and it is not represented. A real registry needs an operational event definition — criteria, source of confirmation, date retained and adjudication procedure. Monitoring and contact with a team may also change care use and the probability that a decompensation is documented, so that more events may reflect better ascertainment, real deterioration, or both; unless the two are separated, a model may learn the surveillance process as much as the physiological trajectory.

- *Retrospective learning* (which trajectories precede events, with what lead time, compared with claims alone and with the T0 assessment): not simulated in the long-term worlds, where detection is a prescribed curve; addressed only by the internal experiment on an injected prodrome (recall 24%, about 19 alerts per true episode at a 5% quota; §3.1).

- *Validation before clinical use* (later period, another territory, silent-mode alerts): not simulated.

- *Escalation and its consequences* (triage, referral to the right level of care, time to assessment, workload of the teams, emergency-department load, unnecessary referrals): not modelled beyond a capacity and a response delay (§3.6); decompensations and deaths are modelled.

A further caveat applies to the registry itself: patients are observed after invitation, assessment, equipment and contact with a team, all of which may change their behaviour and care. A registry built this way learns trajectories under the programme, not the natural history of an untreated disengagement.

The benchmark therefore supports the principle of denser surveillance during disengagement under explicit assumptions, not the predictive value of patterns that remain to be learned, nor the efficacy of a triage model, nor a relief of clinical teams. Escalation would require defining what counts as a sufficiently reliable and urgent pattern, the clinical thresholds and the response circuit, and measuring the load that alerts generate; without them, continuous surveillance may shift the burden to alert teams or acute services instead of relieving professionals. With a positive predictive value of about 5% per alert in the internal experiment, that risk is not theoretical.

5. Conclusion

The main contribution of this paper is a simulation benchmark in which truth is known by construction, showing that, in the mechanisms tested, claims-based evaluation designs misstate the effect of medical disengagement: selecting on hospital reappearance overstated its cost under every mechanism, while the direction of the mortality bias depended on the causal mechanism posited. Within this benchmark, and for type 2 diabetes as a first use case, medical disengagement appears as a loss of life rather than a hidden saving, a consequence of the posited mechanism that real data must still confirm. What the benchmark shows is methodological, and limited to the configurations tested: there, the claims-based designs tested overstated, diluted or postponed a known effect, and across the 14 posited mechanisms tested the upward bias of reappearance-based cohorts persisted while the direction of the mortality bias did not. Synthetic matched worlds make the bias of an evaluation design measurable before any patient is exposed to it, and move the economic argument for vigilance from "how much will it save?" to "how much may it cost?". The headroom computed here answers a formal question — which fixed operating cost would be compatible with a reading threshold under the posited assumptions — and not the practical question of what PREDICARE may cost: it rests on uncalibrated clinical parameters, an unmeasured attenuation of excess risk, undocumented attendance and resumption of care, indirect renal utilities, posited dependency transitions and unknown operating costs. A pre-registered, randomised evaluation of the offer can estimate the effect of the programme as a whole, not, by itself, how that effect divides between its mechanisms. For the intended programme, targeting in claims, a mobile-unit assessment, then monitoring during disengagement, that pilot should randomise the offer and keep every eligible person in the denominator, because, if attendance is selective as posited here, those who come are not those who stay away.

Data and code availability

The generator (predicare-synth v0.19), its configuration files, validity domains, decision gates, provenance registry, 25 architecture decision records, automated tests, experiment scripts and the run record of every result reported (Table 12; Supplementary Table S6) are versioned and reproducible by seed. The SNDS-like tables (20,000 patients, about 10 million rows), the emulated cohorts and the ground-truth directory are available as a separate synthetic dataset. Parameters, calibration status, distributions and implementation details are given in Supplementary Tables S1–S14, and the value of every table by job and seed in a results file of the code deposit. The software is archived on Zenodo (concept DOI

https://doi.org/10.5281/zenodo.23474999; this version, v0.19: https://doi.org/10.5281/zenodo.23475000) under the Apache License 2.0, and the synthetic dataset — SNDS-like tables, emulated cohorts, ground-truth directory, per-seed run records and the supplementary material — as a separate deposit (concept DOI https://doi.org/10.5281/zenodo.23485751; this version https://doi.org/10.5281/zenodo.23485752) under CC-BY-4.0; the Cnam expenditure extraction used for unit costs is included in that deposit under its original ODbL licence. Results of world PE use a remote-monitoring cost of 110 € per equipped month: Tables 10–11 and Figure 6 were re-run with predicare-synth v0.19.2, which reproduces v0.19.1 bit for bit under default settings (archive predicare-synth-v0.19.2.zip, SHA-256 b14994c3ef6cfb461f634d3f07ddc633b45eb8f4fa7f0f44ea49738098edfb45), and the probabilistic analysis B was corrected exactly from its re-run draws; scripts and run records are in version 0.19.2 of the dataset [version DOI to be added at deposit]. No real patient data were used at any stage.

Table 12 - Provenance and status of the results reported: replicated, single-seed or historical; version, replication, population and script (full map in Supplementary Table S6). "Conditional": depends on posited programme parameters (attendance, attenuation, registry curve, programme costs).

Result (where reported)	Status	Version	Replication	Population / estimand	Script (code deposit)
Paired contrasts and claims designs (Abstract; Tables 5–6; Figures 2–4)	Replicated	v0.17	10 seeds × 60,000	E1, E2, E3	`revision/r1_runner.py`, job `base` (exp09, exp14)
Parametric sensitivity, 13 scenarios (§3.2–3.3)	Replicated	v0.19	10 seeds × 60,000 per scenario	E1, E4	`revision/sens_param_r1.py`, jobs `p9`
Cohort emulation and survival analyses (Table 7)	Single seed	v0.14 = v0.19	1 seed × 20,000 (seed 42)	E3	`export_snds_long` → `ciblage_cohortes_snds` → `evaluer_cohortes`
Cost-utility of re-engagement (Table 8; Figure 5)	Replicated	v0.17	10 seeds × 60,000	E4	job `base` (exp10)
IoMT variant PC (Supplementary Table S11; Figure S3)	Replicated; conditional	v0.19	10 seeds × 60,000	E4	jobs `e11`
Cost by population	Replicated	v0.17 / v0.19	10 × 60,000	E2, E4	jobs `base`, `e15`

and perspective (Table 9; Supplementary Table S12)			(payer); 10 × 30,000 (collective)		
Probabilistic analyses A and B (§3.4–3.5; Supplementary Figures S4–S5)	Replicated (300 draws); conditional	v0.19 (B: v0.19.2, 110 €)	300 draws × 20,000	E4	`revision/psa_r1.py`, `revision/psa_analyse_r1.py`
PREDICARE sequence and its estimands (Table 10; Figure 6; §3.5)	Replicated; conditional	v0.19.2 (110 €)	10 seeds × 60,000	E4–E7	jobs `e16`, `e16b`; `analyses/kit_realiste/R1/r1_kit.py`
T0 assessment, capacity, delay; registry ceiling (Table 11; §3.6)	Replicated; conditional	v0.19.2 (110 €; smax: v0.19, 25 €)	10 seeds × 60,000	E4	jobs `e17`, `smax`; `analyses/kit_realiste/R1/r1_kit.py`
Structural experiment (§3.7; Figure 7)	Replicated	v0.18	14 mechanisms × 10 seeds × 60,000	E1, E3, E4	`revision/exp_struct_r1.py`, jobs `s2`
Internal prediction experiment (§3.1)	Replicated; internal	v0.18	10 seeds (daily trunk)	E9	`revision/exp13b_prediction_r1.py`
Evaluation designs (§3.1; Supplementary Table S8)	Replicated	v0.18	10 seeds × 40,000 (daily trunk)	avoided events (ITT)	`revision/exp05b_designs_r1.py`
Historical daily-trunk results (Supplementary Table S9-D)	Historical, not re-run	v0.1–v0.14	1–5 seeds	E8, E9	`experiments/exp02`, `exp06–exp08`, `tools/`

Ethics

The study used exclusively synthetic data; no ethics approval or data-protection authorisation was required. Any real-data use of the targeting script falls under the legal framework of the SNDS and is outside the scope of this paper.

Reporting

The economic evaluation is reported following CHEERS 2022 [36] (checklist in Supplementary Table S4). The simulation model is described following the ISPOR–SMDM good-practice recommendations on model transparency, validation and uncertainty analysis [28,31]; its calibration status is reported in Supplementary Table S2.

The simulation study is reported against the STRESS guidelines [37] (agent-based variant, with the discrete-event item on queues; Supplementary Table S5), and its design follows the ADEMP structure for simulation studies [3].

Supplementary material

Supplementary Tables S1 (parameters, values, sources and epistemic status; S1-C, cost inventory in 2024 euros), S2 (calibration status), S3 (probabilistic sensitivity analyses and implementation details), S4 (CHEERS 2022 checklist), S5 (STRESS checklist), S6 (results map: version, script, command and replication of every result), S7 (structural experiment), S8 (evaluation designs), S9 (detection experiments: internal prediction experiment, registry-ceiling sensitivity and historical campaigns), S10 (estimands of the offer: all eligible versus invited persons), S11 (IoMT variant PC), S12 (collective perspective) and S13 (automated tests), and Supplementary Figures S1 (PREDICARE pathway with cost and health-effect entry points), S2 (calibration in the internal prediction experiment), S3 (registry curve of world PC), S4 and S5 (probabilistic analysis A), together with the results of every table by job and seed, are provided in a separate file and will be deposited with the code.

Funding

Funded by the Twingital-Institute.

Competing interests

Jérôme Vétillard is Vice-President R&D and Chief Product Officer at Qualees (TweenMe), which develops digital-twin and vigilance solutions, including the operational strand of the PREDICARE programme (Sentinelle IA), and is the founder of the Twingital Institute, which carries the PREDICARE doctrinal program. To allow independent scrutiny of results produced under this conflict, the complete code, configuration, run records and synthetic data are deposited openly (Data and code availability), and every result in this paper can be regenerated from them.

Author contributions

Jérôme Vétillard: conceptualisation, methodology, supervision, validation, writing, review and editing.

Use of generative AI

The TweenMe® Platform was used to implement and orchestrate the agent-based simulation components of the PREDICARE demonstrator. The author made and reviewed the design choices, results and text, and takes full responsibility for the content.

Acknowledgements

None

References

References are numbered in order of first citation (Vancouver style). Bibliographic records and DOIs were checked against PubMed and the publishers' or agencies' pages on 16 October 2026; the three references added in this revision (Morris et al., O'Hagan et al. and INSEE) on 17 October 2026.

1. Inspection générale des affaires sociales, Inspection générale des finances. Revue de dépenses relative aux affections de longue durée. Paris: IGAS-IGF; 2024 (erratum July 2025). Available from: https://igas.gouv.fr/sites/igas/files/2025-07/Rapport%20Igas-IGF%20affections%20longue%20dur%C3%A9e%20(revue%20de%20d%C3%A9penses).pdf

2. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758-64. doi:10.1093/aje/kwv254

3. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-102. doi:10.1002/sim.8086

4. Ndumele CE, Rangaswami J, Chow SL, et al. Cardiovascular-kidney-metabolic health: a presidential advisory from the American Heart Association. Circulation. 2023;148(20):1606-35. doi:10.1161/CIR.0000000000001184

5. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int. 2024;105(4S):S117-S314. doi:10.1016/j.kint.2023.10.018

6. Brownrigg JR, Hughes CO, Burleigh D, et al. Microvascular disease and risk of cardiovascular events among individuals with type 2 diabetes: a population-level cohort study. Lancet Diabetes Endocrinol. 2016;4(7):588-97. doi:10.1016/S2213-8587(16)30057-2

7. Yen FS, Wei JC, Shih YH, Hsu CC, Hwu CM. Impact of individual microvascular disease on the risks of macrovascular complications in type 2 diabetes: a nationwide population-based cohort study. Cardiovasc Diabetol. 2023;22(1):109. doi:10.1186/s12933-023-01821-8

8. Giugliano D, Scappaticcio L, Longo M, et al. GLP-1 receptor agonists and cardiorenal outcomes in type 2 diabetes: an updated meta-analysis of eight CVOTs. Cardiovasc Diabetol. 2021;20(1):189. doi:10.1186/s12933-021-01366-8

9. McGuire DK, Shih WJ, Cosentino F, et al. Association of SGLT2 inhibitors with cardiovascular and kidney outcomes in patients with type 2 diabetes: a meta-analysis. JAMA Cardiol. 2021;6(2):148-58. doi:10.1001/jamacardio.2020.4511

10. Plateforme des données de santé (Health Data Hub). Données synthétiques – top diabète [dataset]. data.gouv.fr; 2024. Licence Ouverte 2.0. Available from: https://www.data.gouv.fr/datasets/donnees-synthetiques-top-diabete

11. Caisse nationale de l'Assurance Maladie. Pathologies : dépenses remboursées affectées à chaque pathologie, 2015–2024 [dataset]. data.gouv.fr; release of 6 July 2026 (2024 vintage; full CSV, SHA-256 8b05ec57…6816505). Available from: https://www.data.gouv.fr/datasets/pathologies-depenses-remboursees-affectees-a-chaque-pathologie

12. Institut national de la statistique et des études économiques (INSEE). Indice des prix à la consommation – ensemble, base 2015, moyennes annuelles. Paris: INSEE. Available from: https://www.insee.fr/fr/statistiques/6036866

13. Agence de la biomédecine. Rapport annuel REIN 2024 (Réseau épidémiologie et information en néphrologie). Saint-Denis: ABM; 2026. Available from: https://www.agence-biomedecine.fr/en/observatoire-de-la-maladie-renale-chronique-reports

14. Chaudhry SI, Wang Y, Concato J, Gill TM, Krumholz HM. Patterns of weight change preceding hospitalization for heart failure. Circulation. 2007;116(14):1549-54. doi:10.1161/CIRCULATIONAHA.107.690768

15. Boehmer JP, Hariharan R, Devecchi FG, et al. A multisensor algorithm predicts heart failure events in patients with implanted devices: results from the MultiSENSE study. JACC Heart Fail. 2017;5(3):216-25. doi:10.1016/j.jchf.2016.12.011

16. Clarke P, Gray A, Holman R. Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62). Med Decis Making. 2002;22(4):340-9. doi:10.1177/0272989X0202200412

17. Stout NK, Goldie SJ. Keeping the noise down: common random numbers for disease simulation modeling. Health Care Manag Sci. 2008;11(4):399-406. doi:10.1007/s10729-008-9067-6

18. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444-55. doi:10.2307/2291629

19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html

20. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107. doi:10.1002/sim.3697

21. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-81. doi:10.2307/2281868

22. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Stat. 1978;5(3):141-50.

23. Hagenmeyer EG, Koltermann KC, Dippel FW, Schädlich PK. Health economic evaluations comparing insulin glargine with NPH insulin in patients with type 1 diabetes: a systematic review. Cost Eff Resour Alloc. 2011;9:15. doi:10.1186/1478-7547-9-15

24. Beaudet A, Clegg J, Thuresson PO, Lloyd A, McEwan P. Review of utility values for economic modeling in type 2 diabetes. Value Health. 2014;17(4):462-70. doi:10.1016/j.jval.2014.03.003

25. Haute Autorité de santé. Choix méthodologiques pour l'évaluation économique à la HAS [Choices in methods for economic evaluation]. Saint-Denis La Plaine: HAS; 2020. Available from: https://has-sante.fr/upload/docs/application/pdf/2020-11/methodological_guidance_2020_-choices_in_methods_for_economic_evaluation.pdf

26. Boneschi S, Miron de l'Espinay A. Aides à l'autonomie des personnes âgées : qui paie quoi ? L'apport du modèle Autonomix – résultats 2019. Les Dossiers de la DREES n° 99. Paris: DREES; 2022. Available from: https://drees.solidarites-sante.gouv.fr/sites/default/files/2022-07/dd99.pdf

27. Besnard X, Brunel M, Couvert N, Roy D. Les proches aidants des seniors et leur ressenti sur l'aide apportée – résultats des enquêtes « CARE » auprès des aidants (2015-2016). Les Dossiers de la DREES n° 45. Paris: DREES; 2019. Available from: https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-07/dd45.pdf

28. Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making. 2012;32(5):722-32. doi:10.1177/0272989X12458348

29. O'Hagan A, Stevenson M, Madan J. Monte Carlo probabilistic sensitivity analysis for patient level simulation models: efficient estimation of mean and variance using ANOVA. Health Econ. 2007;16(10):1009-23. doi:10.1002/hec.1199

30. Strong M, Oakley JE, Brennan A. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a nonparametric regression approach. Med Decis Making. 2014;34(3):311-26. doi:10.1177/0272989X13505910

31. Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making. 2012;32(5):733-43. doi:10.1177/0272989X12454579

32. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58(1):21-9. doi:10.1111/j.0006-341x.2002.00021.x

33. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182-91. doi:10.1016/j.cct.2006.05.007

34. Walonoski J, Kramer M, Nichols J, et al. 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(3):230-8. doi:10.1093/jamia/ocx079

35. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219-26. doi:10.1016/j.csda.2013.10.018

36. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. BMJ. 2022;376:e067975. doi:10.1136/bmj-2021-067975

37. Monks T, Currie CSM, Onggo BS, Robinson S, Kunc M, Taylor SJE. Strengthening the reporting of empirical simulation studies: introducing the STRESS guidelines. J Simul. 2019;13(1):55-67. doi:10.1080/17477778.2018.1442155