Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid recommendation of healthcare interventions for personalised patient engagement: a leakage-free offline evaluation on synthetic electronic health records

Sefunmi Ajayi
University of Ulster
Tertsegha Joseph Anande
t.anande@ulster.ac.uk

University of Ulster

Research Article

Keywords: Health recommender systems, Hybrid filtering, Patient engagement, Synthetic electronic health records, Offline evaluation, Personalised healthcare

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11529962/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid recommendation of healthcare interventions for personalised patient engagement: a leakage-free offline evaluation on synthetic electronic health records

Sefunmi Ajayi[1] and Tertsegha Joseph Anande[1*]

[1]Department of Computing, Ulster University, Birmingham, United Kingdom.

*Corresponding author(s). E-mail(s): t.anande@ulster.ac.uk;
Contributing authors: sefunmiajayi4@gmail.com;

Abstract

Digital health platforms often deliver the same content to every patient, which limits engagement. Recommender systems could tailor care plans, medications and procedures to the individual, but evaluations in health settings often lack strong baselines, select hyperparameters on test data, or rely on a single random split. We built a catalogue of 300 recommendable interventions from the April 2020 Synthea sample of 1,171 synthetic patients and compared nine recommenders: random and popularity baselines, content-based filtering on clinical reason codes (CBF), CBF with data-derived item profiles (CBF-D), user-based nearest-neighbour collaborative filtering (kNN), the linear autoencoder EASE, and three weighted hybrids. All hyperparameters were tuned on validation data carved from training only. We report patient-level bootstrap confidence intervals, paired tests, random and temporal splits, ten split seeds, cohort sizes from 113 to 1,171 patients, subgroup performance and a cold-start scenario. The CBF + kNN hybrid outperformed both of its components (NDCG@10 0.692 versus 0.643 for kNN and 0.319 for CBF) with a selected content weight of 0.3. However, the tuned EASE baseline scored higher (0.754), and the best model combined CBF-D with EASE (0.787, 95% CI 0.773–0.802). Under the temporal split, every hybrid outperformed every single-signal model, including EASE. A popularity baseline reached a hit ratio of 0.682 while covering only 9.3% of the catalogue. Hybrid recommendation is a sound basis for personalised engagement, but strong baselines and leakage-free tuning are needed to measure its benefit. Validation on real records with engagement outcomes is the next step.

Keywords: Health recommender systems, Hybrid filtering, Patient engagement, Synthetic electronic health records, Offline evaluation, Personalised healthcare

1 Introduction

Patient engagement, the extent to which people take an active part in their own care, is consistently associated with better adherence, outcomes and care experiences [1, 2]. Digital health platforms have widened the channels through which engagement can be supported, yet many still deliver generic content that ignores the individual patient's conditions, history and preferences [3]. Stakeholders repeatedly name personalisation, relevance and trust as the conditions under which patients engage with digital tools [3, 4], and artificial intelligence (AI) is increasingly proposed as the means to provide them, from engagement estimation in telehealth [5] and patient-centred clinical decision support [6] to personalised treatment platforms [7] and generative conversational agents [8].

Recommender systems are the most established AI technique for personalisation. Health recommender systems (HRS) adapt the content-based, collaborative and hybrid approaches of e-commerce and media to suggest health information, behaviours or interventions [9–11]. Systematic reviews, however, describe a field in which evaluation is heterogeneous and frequently weak: many systems are assessed on small samples, without comparison to simple baselines, and with little attention to reproducibility [12, 13]. The wider recommender-systems literature has shown that such practices can make new methods appear better than they are. Well-tuned simple baselines often match or beat sophisticated models [14, 15]. Tuning on test data, random rather than temporal splits and sampled metrics inflate results [16–18]. Popularity effects can masquerade as personalisation [19, 20].

This article develops an MSc project that designed a hybrid recommender for personalised patient engagement on Synthea synthetic electronic health records (EHRs) [21] and embedded it in a prototype platform, *Engage*. The project made three design choices that we retain. It defined recommendable interventions by pooling care plans, medications and procedures into a single catalogue. It combined content-based and collaborative scores in a weighted hybrid. It surfaced each recommendation with its clinical rationale. Its evaluation on 113 patients reported that the hybrid outperformed both component models with a content weight of $\alpha = 0.3$. That evaluation had no popularity or random baseline, used one random split without uncertainty estimates, and selected α on scores that coincided with the held-out test results. Here we re-implement the pipeline on a larger, versioned public Synthea release and subject it to an evaluation designed to the standards of current recommender-systems research.

Our contributions are:

1. A reproducible method, with released code, for turning Synthea-style EHR tables into a catalogue of recommendable interventions and a patient–intervention interaction matrix.

2. A leakage-free comparison of nine recommenders, including popularity and the strong linear baseline EASE [22], under random and temporal splits, with bootstrap confidence intervals and paired significance tests.
3. An enhanced content model (CBF-D) that derives item profiles from the conditions of patients who received each intervention. It addresses the incomplete reason coding that limited the original content-based model.
4. Robustness evidence on split seeds, cohort size (including the original $n = 113$), demographic subgroups and cold-start recommendation for newly enrolled patients, together with a description of how the findings inform the Engage prototype.

2 Related work

Health recommender systems.

Wiesner and Pfeifer framed HRS as systems whose recommended items are medically relevant and whose users may be patients or professionals [9]. Subsequent work distinguished health-aware recommendation, which must weigh benefit and risk rather than click-through [10]. It also catalogued applications in health promotion, lifestyle change, medication and care-pathway recommendation [11–13]. De Croon et al. reviewed 73 HRS studies and found that most evaluated only algorithmic accuracy, often on small or proprietary data sets, with few user studies [13]. Equity and trust are recurring concerns: AI-enabled engagement tools can widen disparities if their data under-represent groups [23, 24], and patients' willingness to act on AI advice depends on transparency and explanation [4, 25, 26].

Content-based, collaborative and hybrid recommendation.

Content-based filtering (CBF) matches item descriptors to a user profile [27]. It handles new items but depends on the quality of item metadata. Collaborative filtering (CF) exploits co-occurrence across users, classically through user-based neighbourhoods [28]. More recently, linear item–item models such as EASE learn the full item–item weight matrix in closed form and are competitive with deep models [14, 22]. Hybrids combine the two to offset each other's weaknesses, most simply by weighted score blending [29].

Offline evaluation.

Ranking metrics such as precision, recall, normalised discounted cumulative gain (NDCG) [30] and hit ratio are standard [31], but their values depend heavily on protocol. Random per-user holdout lets models learn from interactions that occurred after the ones they must predict, whereas temporal splits better reflect deployment [16, 17]. Sampled-negative evaluation can reverse model rankings, so full-catalogue ranking is preferred [18]. Reproducibility studies recommend tuned simple baselines, held-out validation for hyperparameter selection, and repeated runs [14, 15].

Synthetic EHRs.

Synthea generates longitudinal patient records from disease modules informed by clinical guidelines and population statistics [21]. Synthetic data avoid privacy constraints

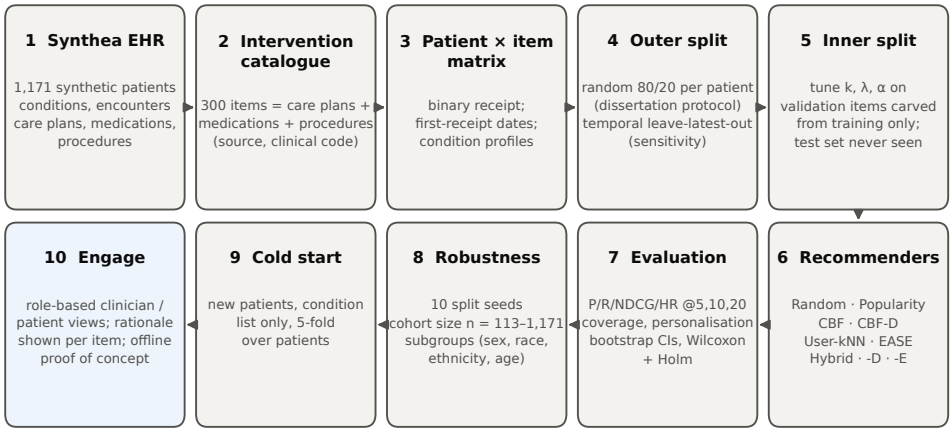


Fig. 1 Study pipeline. The intervention catalogue, weighted hybrid and Engage interface follow the original MSc design; steps 4–5 and 7–9 add the evaluation safeguards described in Section 3.

and support reproducible method development [32]. However, Synthea's realism is limited: it reproduces many population-level patterns, but performs poorly on clinical quality measures that depend on care variation and adherence [33]. Results on Synthea are therefore best read as evidence about methods and pipelines rather than about clinical effectiveness.

3 Methods

3.1 Study design

The study follows design science research [34]: an artefact (an intervention recommender and its prototype interface) is designed to address a practical problem and evaluated rigorously. Figure 1 summarises the pipeline. All analyses are implemented in a single script. After a first run, two changes were made: the neighbourhood-size grid was widened because the selected value lay at its upper edge, and a CBF-D + EASE hybrid (Hybrid-E) was added because EASE was the strongest single model. Both changes are reported, and neither involved the test data in model selection.

3.2 Data

We used the Synthea sample data set in CSV format, April 2020 release [35]. It contains 1,171 synthetic patients resident in Massachusetts with records to 28 April 2020. The release is fixed and publicly downloadable, so every number reported here can be regenerated. Patients' age was computed at death or at the latest encounter date in the data set. The original project used a smaller Synthea extract of 113 patients; Table 1 compares the two.

3.3 Intervention catalogue and interactions

Synthea does not label recommendable items. As in the original project, we pooled three record types as interventions: care plans, medications and procedures. Each unique (source type, clinical code) pair defines one catalogue item. Immunisations were excluded because they follow a fixed schedule. Each item carries the set of reason codes recorded with it, where present. The interaction matrix $\mathbf{X} \in \{0,1\}^{m \times n}$ has $x_{ij} = 1$ if patient i ever received intervention j, and we keep the date of first receipt for the temporal split. Each patient's condition profile $\mathbf{c}_i$ is a multi-hot vector over condition codes.

3.4 Recommendation models

All models score every catalogue item a patient has not received in the training data, and the top K are recommended. Scores are ranked against the full catalogue; no negative sampling is used [18].

Baselines.

Random draws uniform scores. *Popularity* ranks items by the number of training patients who received them, so every patient receives the same list.

Content-based filtering (CBF).

This is the dissertation's design. A patient is represented by their condition profile concatenated with the proportions of their training interventions from each source type, $\mathbf{p}_i = [\mathbf{c}_i; \mathbf{s}_i]$. An item is represented by its reason codes in the same condition vocabulary concatenated with a one-hot source indicator, $\mathbf{r}_j = [\mathbf{g}_j; \mathbf{e}_j]$. The score is cosine similarity,

$$\hat{y}_{ij}^{\text{CBF}} = \frac{\mathbf{p}_i^\top \mathbf{r}_j}{\|\mathbf{p}_i\| \, \|\mathbf{r}_j\|}. \tag{1}$$

CBF with data-derived item profiles (CBF-D).

Items without a reason code have no clinical content under CBF. CBF-D instead describes each intervention by the inverse-document-frequency-weighted mean condition profile of the training patients who received it,

$$\mathbf{q}_j = \frac{\sum_u x_{uj}^{\text{tr}} \, (\mathbf{c}_u \odot \mathbf{w})}{\sum_u x_{uj}^{\text{tr}}}, \qquad w_c = \log \frac{1+m}{1+\text{df}_c} + 1, \tag{2}$$

and scores $\hat{y}_{ij}^{\text{CBF-D}} = \cos(\mathbf{c}_i \odot \mathbf{w}, \, \mathbf{q}_j)$. The model still matches a patient's conditions to item content, but the content is learned from training data rather than taken from incomplete coding.

User-based collaborative filtering (kNN).

This follows the dissertation. Patient similarity is the cosine of training interaction rows, and a patient's score for item j is the similarity-weighted average over the k

most similar patients $N_k(i)$ [28]:

$$\hat{y}_{ij}^{\mathrm{kNN}} = \frac{\sum_{u \in N_k(i)} \mathrm{sim}(i,u)\, x_{uj}^{\mathrm{tr}}}{\sum_{u \in N_k(i)} |\mathrm{sim}(i,u)|}. \tag{3}$$

EASE.

As a strong collaborative baseline, EASE learns an item–item weight matrix with a zero diagonal in closed form, $\hat{\mathbf{B}} = \mathbf{I} - \mathbf{P}\,\mathrm{diagMat}(\mathbf{1} \oslash \mathrm{diag}(\mathbf{P}))$ with $\mathbf{P} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$, and scores $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ [22].

Weighted hybrids.

Following the dissertation, content and collaborative scores are min–max scaled over each patient's candidate items and blended [29]:

$$\hat{y}_{ij}^{\mathrm{hyb}} = \alpha\, \tilde{y}_{ij}^{\mathrm{content}} + (1-\alpha)\, \tilde{y}_{ij}^{\mathrm{collab}}, \qquad 0 \le \alpha \le 1. \tag{4}$$

Hybrid combines CBF with kNN, as in the dissertation. *Hybrid-D* combines CBF-D with kNN, and *Hybrid-E* combines CBF-D with EASE.

3.5 Evaluation protocol

Outer splits.

The primary protocol reproduces the dissertation's random split: for every patient with at least five interventions, 20% of their interventions (at least one) are withheld at random as test items. Patients with fewer than five interventions remain in training only. As a sensitivity analysis, a temporal leave-latest-out split withholds the most recent 20% of each patient's interventions by first-receipt date [17]. Items first received on the cutoff date are also withheld. Only conditions with onset before the cutoff enter that patient's profile.

Leakage-free tuning.

Hyperparameters are selected on an inner split constructed by applying the same procedure to the training data. Grids are $k \in \{5, 10, 15, 20, 30, 50, 75, 100, 150\}$, $\lambda \in \{10, 50, 100, 250, 500, 1000\}$ and $\alpha \in \{0, 0.1, \ldots, 1\}$, with NDCG@10 as the criterion. Models are then refitted on the full training data and scored once on the test items, which are never used for selection.

Metrics.

For each evaluated patient with relevant test set T_i and top-K list L_i^K we compute precision $|L_i^K \cap T_i|/K$, recall $|L_i^K \cap T_i|/|T_i|$, hit ratio $\mathbb{1}[|L_i^K \cap T_i| > 0]$ and NDCG@K with binary relevance [30], at $K \in \{5, 10, 20\}$, then average over patients. Beyond accuracy, catalogue coverage is the share of items appearing in any top-10 list. Personalisation is one minus the mean pairwise Jaccard overlap of top-10 lists [20].

Statistics.

We report 95% confidence intervals from 2,000 patient-level bootstrap resamples. Models are compared with paired Wilcoxon signed-rank tests on per-patient NDCG@10, with Holm adjustment within each family of comparisons. NDCG@10 is the primary end point; other metrics are secondary.

3.6 Robustness analyses

Split seeds: the random protocol, including inner tuning, was repeated with ten further seeds. *Cohort size:* to connect with the original 113-patient study and to examine data scarcity, the full pipeline was rerun on 20 random patient subsamples each of size 113, 250 and 500. *Subgroups:* Hybrid and Hybrid-D NDCG@10 were stratified by recorded sex, race, ethnicity and age band to screen for performance disparities [19, 24]. *Cold start:* the Engage prototype onboards new patients who have no intervention history. We simulated this with five folds over patients. Each held-out patient was scored using only their condition list. The comparators were popularity, CBF (with the training cohort's mean source mix), CBF-D, and a condition-based nearest-neighbour model (Condition-kNN), which averages the interventions of the k training patients with the most similar IDF-weighted condition profiles.

3.7 Prototype

The original project implemented the recommender in *Engage*, a self-contained offline web prototype with role-based views. Care-team users see a ranked list of interventions for each patient, each annotated with its clinical rationale (the matched condition or reason code), and can compare the underlying engines for governance purposes. Patients see plain-language suggestions they can act on and track. Newly enrolled patients receive cold-start recommendations. The prototype stores data only in the local browser and has no authentication or live EHR connection. It is described here as a design demonstration and was not part of the quantitative evaluation.

3.8 Implementation

Analyses were implemented in Python 3.14 with NumPy, pandas, SciPy and Matplotlib. The full pipeline (all experiments) runs in about one minute on a laptop CPU. Code, the data checksum (SHA-256 `4194b18c...cd38a`) and all result files are provided as Online Resource 2.

3.9 Ethics

The study used only fully synthetic, publicly released records and involved no human participants or personal data, so ethical approval was not required.

Table 1 Cohort, intervention catalogue and evaluation splits. The dissertation column reports the values stated in the original MSc study (a different, smaller Synthea extract) for comparison.

Characteristic	This study[a]	Dissertation
Patients	1,171	113
Female / male	609 / 562	60 / 53
Deceased	171	NR
Age, mean (SD), years	41.6 (24.3)	48.2 (27.2)
White / Black / Asian / other	965 / 101 / 90 / 15	99 / 7 / 6 / 1
Hispanic ethnicity	113	NR
Encounters	53,346	7,210
Condition records (distinct codes)	8,376 (129)	4,748 (201)
Source records: care plans	3,483	334
medications	42,989	8,180
procedures	34,981	20,527
Catalogue items (total)	300	418
care plans / medications / procedures	32 / 131 / 137	31 / 150 / 237
with $\geq$1 reason code	180 (60.0%)	208 (49.8%)
Patient–item interactions	17,985	4,307[b]
per patient, mean (median)	15.4 (11)	38.1 (35)
matrix density	5.1%	7.3%[c]
Random split: train / test interactions	14,500 / 3,485	3,456 / 851
patients evaluated	1,004	109
Temporal split: train / test interactions	13,798 / 4,187	–
patients evaluated	1,004	–

NR, not reported. [a]Synthea sample data, April 2020 release (CSV). [b]Train plus test interactions. [c]As reported; 4,307 interactions over 113 × 418 cells imply a density of 9.1%.

4 Results

4.1 Cohort and catalogue

Table 1 describes the cohort. The 300-item catalogue comprised 32 care plans, 131 medications and 137 procedures; 180 items (60.0%) carried at least one reason code, ranging from 58 of 131 medications to 27 of 32 care plans (Fig. 2b). Patients had received a median of 11 distinct interventions (mean 15.4), giving a matrix density of 5.1%. Item popularity was highly skewed: medication reconciliation was recorded for 98.4% of patients and 9 items reached at least 20%, whereas 211 of the 300 items reached fewer than 5% (Fig. 2c). The random split withheld 3,485 interactions from 1,004 patients; the temporal split withheld 4,187 from the same 1,004 patients.

4.2 Hyperparameter selection

On inner validation data, the CBF + kNN hybrid peaked at $\alpha = 0.3$ (validation NDCG@10 0.523, against 0.485 for kNN alone at $\alpha = 0$ and 0.259 for CBF alone at $\alpha = 1$; Fig. 3). This reproduces the weight reported in the dissertation, now selected without access to test data. The hybrids using data-derived content preferred more content weight ($\alpha = 0.5$ for both Hybrid-D and Hybrid-E). Across ten further split seeds the selected values were stable (α 0.3–0.4 for Hybrid; Table S3).

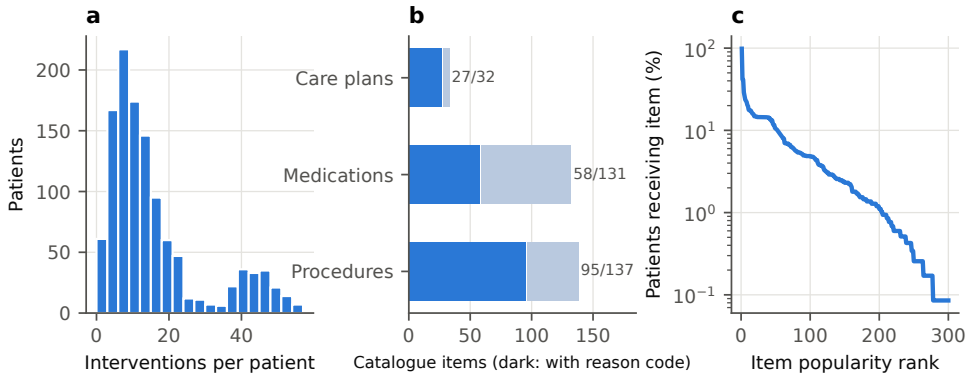


Fig. 2 Cohort and catalogue. **a** Distinct interventions per patient. **b** Catalogue items by source, split by whether a reason code is recorded. **c** Share of patients receiving each item, ranked (log scale).

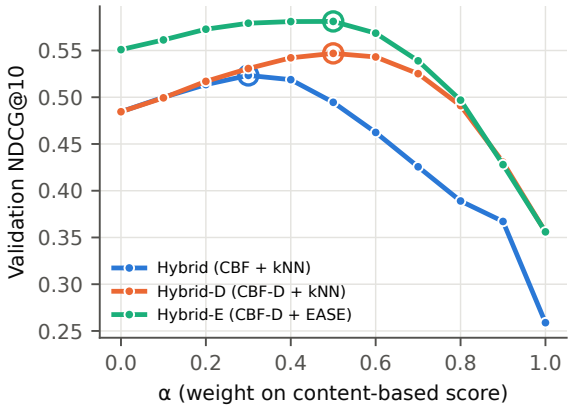


Fig. 3 Inner-validation NDCG@10 as a function of the content weight α for the three hybrids (random split, seed 2026). Open circles mark the selected values. $\alpha = 0$ is the collaborative model alone and $\alpha = 1$ the content model alone.

4.3 Model comparison

Table 2 and Fig. 4 give top-10 performance. Under the primary random split, the dissertation's Hybrid reached NDCG@10 0.692 (95% CI 0.675–0.708). This was higher than both of its components, kNN (0.643) and CBF (0.319), with paired differences of +0.049 and +0.372 (both $p_{\text{Holm}} < 0.001$; Table S5), confirming the original study's central finding. CBF performed no better than popularity on NDCG@10 (0.319 versus 0.314). Replacing reason codes with data-derived profiles raised content-only NDCG@10 to 0.430 and Hybrid-D to 0.724.

The tuned EASE baseline, which the original study did not include, outperformed the dissertation's Hybrid (0.754 versus 0.692; difference −0.062, $p_{\text{Holm}} < 0.001$). Adding CBF-D to EASE improved it further: Hybrid-E achieved the best scores on every top-10 accuracy metric (NDCG@10 0.787, 95% CI 0.773–0.802; P@10 0.284;

Table 2 Top-10 recommendation performance on held-out interventions. Values are means over evaluated patients; NDCG@10 is shown with its 95% patient-level bootstrap confidence interval. Coverage is the share of the 300-item catalogue appearing in any top-10 list. Hit ratio, results at $K = 5$ and $K = 20$, and personalisation for both splits are in Online Resource 1 (Tables S1–S2).

	Random split (primary; $n = 1{,}004$)				Temporal split ($n = 1{,}004$)		
Model	P@10	R@10	NDCG@10	Cov.	R@10	NDCG@10	Cov.
Random	0.011	0.030	0.017 (0.013–0.021)	1.000	0.030	0.024 (0.019–0.028)	1.000
Popularity	0.100	0.368	0.314 (0.294–0.332)	0.093	0.222	0.186 (0.171–0.203)	0.110
CBF	0.158	0.469	0.319 (0.305–0.334)	0.773	0.376	0.283 (0.267–0.301)	0.727
CBF-D	0.175	0.604	0.430 (0.412–0.449)	0.990	0.450	0.339 (0.320–0.359)	0.910
User-kNN CF	0.246	0.733	0.643 (0.625–0.662)	0.700	0.438	0.357 (0.338–0.376)	0.427
EASE	0.278	0.827	0.754 (0.737–0.769)	0.747	0.461	0.374 (0.354–0.392)	0.477
Hybrid	0.257	0.783	0.692 (0.675–0.708)	0.877	0.524	0.421 (0.400–0.441)	0.763
Hybrid-D	0.266	0.812	0.724 (0.708–0.740)	0.973	0.561	0.456 (0.437–0.475)	0.907
Hybrid-E	0.284	0.854	0.787 (0.773–0.802)	0.987	0.577	0.471 (0.451–0.490)	0.893

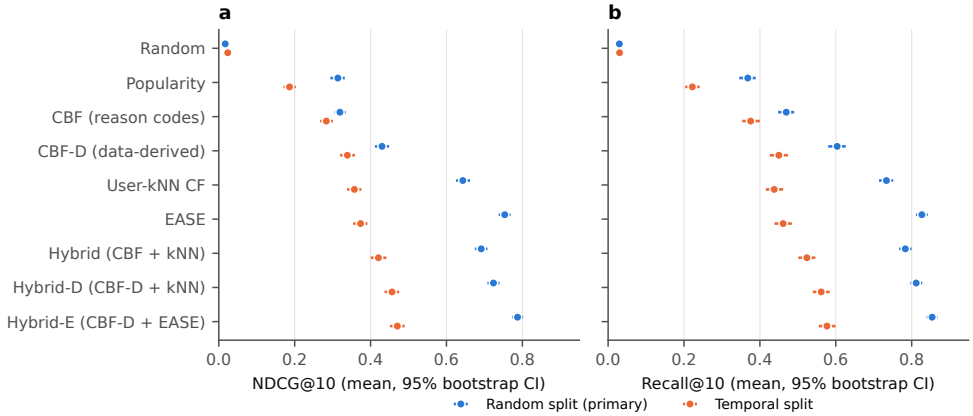


Fig. 4 Top-10 accuracy of all nine models under the random (primary) and temporal splits. Points are means over evaluated patients; bars are 95% patient-level bootstrap confidence intervals.

R@10 0.854; difference from EASE +0.034, $p_{\text{Holm}} < 0.001$). Hybrid-E was also best on every metric at $K = 5$ and $K = 20$ (Table S1).

Hit ratio separated models poorly. Popularity, which gives every patient the same list, achieved HR@10 of 0.682, and all collaborative and hybrid models exceeded 0.93. The beyond-accuracy metrics showed what popularity sacrifices: it covered 9.3% of the catalogue, against 87.7% for Hybrid and 98.7% for Hybrid-E, and its lists were far less individualised (personalisation 0.448 versus 0.886 and 0.935).

4.4 Temporal split

Predicting each patient's most recent interventions from their earlier history was harder for every model (Table 2). It also changed the ranking. Collaborative signals weakened most: kNN fell to NDCG@10 0.357 and EASE to 0.374, with coverage of

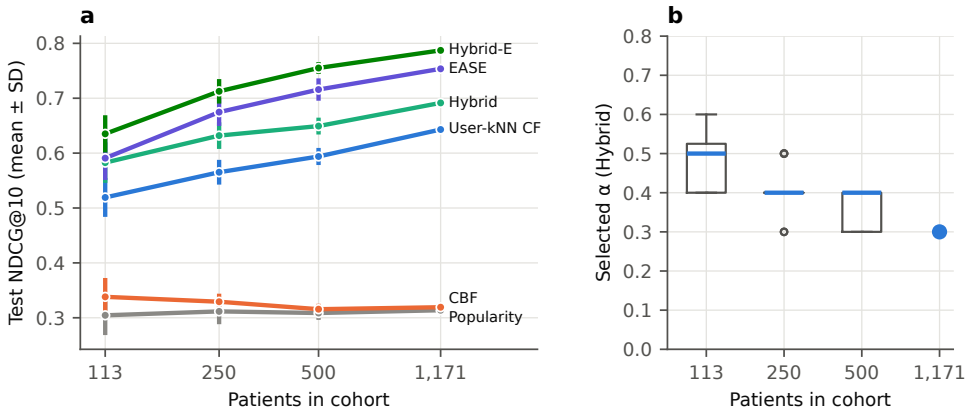


Fig. 5 Effect of cohort size (random split). **a** Test NDCG@10, mean ± SD over 20 subsamples per size; $n = 1{,}171$ is the full cohort. **b** Content weight α selected on inner validation data for the CBF + kNN hybrid.

42.7% and 47.7%. Content-based models degraded less. Every hybrid now outperformed every single-signal model, including EASE: Hybrid 0.421 (0.400–0.441; +0.047 versus EASE, $p_{\text{Holm}} < 0.001$), Hybrid-D 0.456 and Hybrid-E 0.471 (0.451–0.490). With temporal selection, the Hybrid content weight rose to $\alpha = 0.4$ (Table S3).

4.5 Robustness

Across ten further random-split seeds, NDCG@10 varied little: SDs were 0.011 or less for every model, and the ranking of mean scores matched the primary analysis (Table S4). Performance rose with cohort size for all collaborative and hybrid models (Fig. 5a; Table S6). At the dissertation's size of 113 patients, Hybrid (mean NDCG@10 0.583) clearly beat kNN (0.519) and was close to EASE (0.591), while Hybrid-E remained best (0.635). The selected content weight fell as data grew, from a mean α of 0.49 at $n = 113$ to 0.30 at $n = 1{,}171$ (Fig. 5b). Content information mattered most when collaborative evidence was scarce.

Subgroups.

Hybrid NDCG@10 was similar across recorded sex (female 0.702, 95% CI 0.680–0.723; male 0.679, 0.652–0.705), race, ethnicity and age band (Fig. 6; Table S7). Every subgroup's confidence interval included the overall mean of 0.692. The small "other/native" group ($n = 15$) had a wide interval, so no inference can be drawn for it.

4.6 Cold start

For newly enrolled patients scored on their condition list alone, Condition-kNN was most accurate (NDCG@10 0.798, 95% CI 0.788–0.808; P@10 0.612; Table 3). CBF-D (0.566) outperformed popularity (0.538), whereas reason-code CBF did not (0.471). Popularity achieved a near-perfect hit ratio (0.996) because almost every patient

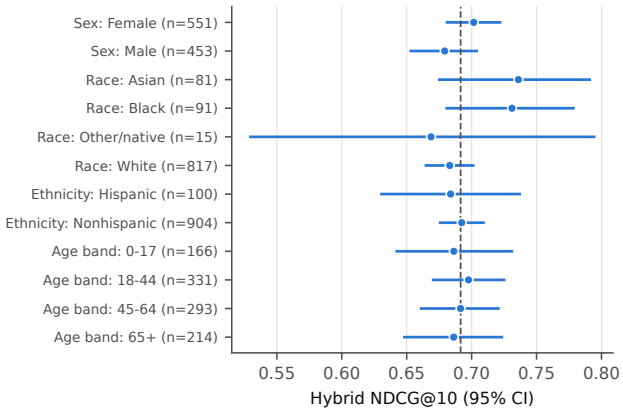


Fig. 6 Hybrid NDCG@10 by subgroup (random split), with 95% bootstrap confidence intervals. The dashed line is the overall mean.

Table 3 Cold-start performance for newly enrolled patients with no intervention history ($n = 1{,}167$, five folds over patients). Only the patient's condition list is used; relevance is every intervention the patient went on to receive. Values are means; NDCG@10 with 95% bootstrap CI.

Model	P@10	R@10	NDCG@10	HR@10
Popularity	0.346	0.370	0.538 (0.528–0.549)	0.996
CBF (reason codes)	0.417	0.334	0.471 (0.457–0.485)	0.930
CBF-D (data-derived profiles)	0.468	0.442	0.566 (0.553–0.580)	0.969
Condition-kNN	0.612	0.570	0.798 (0.788–0.808)	0.985

received at least one of the commonest interventions, again showing that hit ratio alone is not informative here.

5 Discussion

5.1 Principal findings

This re-evaluation confirms the central result of the original MSc study and puts it in context. With hyperparameters selected on validation data only, the weighted hybrid of content-based and user-based collaborative filtering outperformed both components on a cohort ten times larger. The selected content weight was the same $\alpha = 0.3$. The magnitude of the benefit, however, depended on what the hybrid was compared with and how the data were split. A tuned linear collaborative model (EASE) outperformed the original hybrid under the random split. Hybridisation still helped: adding data-derived content to EASE produced the best model overall. Under the more deployment-like temporal split, all three hybrids outperformed all single-signal models. Predicting a patient's next interventions from their earlier history evidently needs both signals: patients who are similar so far, and the clinical conditions that drive what comes next.

Two further results matter for the original design. First, reason-code CBF was no better than popularity, because 40% of catalogue items, mostly medications, carry

no reason code. Learning item content from the conditions of their recipients (CBF-D) removed most of this deficit without external ontologies. This is a lightweight alternative to the SNOMED CT mapping proposed in the dissertation. Second, the relationship between cohort size and selected α suggests a practical rule: platforms with few patients should weight content more heavily and recalibrate α as their data grow. Engage already exposes α as a configurable setting for this purpose.

5.2 Methodological lessons

Several features of typical HRS evaluations would have misled us here. Hit ratio was uninformative: a non-personalised popularity list hit at least one relevant intervention for 68% of patients in the main evaluation and for nearly all new patients in the cold-start test. The original study's hit ratios of 0.83–0.99 should be read in that light. Without a strong baseline, the hybrid would have appeared to be the best available model. Tuning on test data inflates every tuned model, and the original tuning curve coincided with its test results. Patient-level confidence intervals, paired tests and repeated splits cost little and showed that the main differences here are large relative to sampling variation. These lessons echo reproducibility studies in mainstream recommender research [14–16] and the call for more rigorous HRS evaluation [13].

5.3 Implications for engagement platforms

For a platform such as Engage, the results support three design choices. The first is a hybrid default engine, with Hybrid-E preferred where a linear item model can be maintained. The second is a dedicated cold-start path based on condition similarity for newly enrolled patients. The third is per-item rationales. CBF-D's item profiles also yield an explanation: the conditions most characteristic of the patients who received an intervention. This supports the transparency that patients and clinicians expect of AI-enabled engagement tools [4, 25, 36]. Monitoring beyond-accuracy metrics guards against popularity collapse, in which every patient is shown the same routine items, a failure that accuracy metrics can hide.

5.4 Limitations

First, all data are synthetic. Synthea's disease modules generate interventions from guideline-like rules, so models partly learn the generator's logic, and real records contain variation, non-adherence and documentation gaps that Synthea under-represents [32, 33]. Absolute scores should not be expected to transfer. Second, relevance is defined as *received* interventions, an imperfect proxy for interventions that are clinically appropriate or that patients would engage with. Neither engagement nor outcomes were measured, so the study shows recommendation accuracy, not improved engagement. Third, the subgroup analysis reflects the demographic make-up of a synthetic Massachusetts population, with small minority groups, and cannot establish fairness in real populations. Fourth, the temporal split is per patient: collaborative models may learn from other patients' interactions dated after a given patient's cutoff. Fifth, the Engage prototype was not evaluated with users, and it lacks authentication, a secure back end and EHR integration.

5.5 Future work

Validation on de-identified real EHR data with appropriate governance is the essential next step, ideally with an outcome closer to engagement, such as uptake of recommended care-plan activities. Clinical safety constraints (contraindications, guideline conformance) should be added before any patient-facing use, consistent with health-aware recommendation [10]. Finally, a user study with clinicians and patients should assess the usability, trust and actionability of Engage's recommendations and explanations.

6 Conclusion

Hybrid recommendation of care plans, medications and procedures from EHR data is a sound basis for personalising digital patient engagement. Under leakage-free evaluation, a weighted hybrid beat its components; adding data-derived content to a strong linear collaborative model gave the best results; and hybridisation mattered most in the realistic temporal setting and for small cohorts. These conclusions rest on strong baselines, validation-only tuning and uncertainty estimates, which we recommend as minimum practice for health recommender evaluation. Confirmation on real patient records, with engagement outcomes, should come before clinical deployment.

Supplementary information. Online Resource 1 (PDF) contains Tables S1–S7. Online Resource 2 (ZIP) contains the analysis code, the data download script and checksum, and all result files.

Acknowledgements. The authors thank the MITRE Corporation and the Synthea project for making synthetic patient data openly available.

Declarations

- **Funding:** No funding was received for this study.
- **Competing interests:** The authors have no competing interests to declare.
- **Ethics approval and consent to participate:** Not applicable; the study used only publicly available synthetic data and involved no human participants.
- **Consent for publication:** Not applicable.
- **Data availability:** The Synthea sample data (April 2020 CSV release) are publicly available from the Synthea project under the Apache License 2.0 [35].
- **Code availability:** All code needed to reproduce the analyses, tables and figures is provided in Online Resource 2.
- **Use of generative AI:** Generative AI (Claude, Anthropic) assisted with re-implementing the analysis code, producing figures and editing the manuscript. The authors specified the analyses, reviewed all code and text, verified every reported number against the generated result files, and take full responsibility for the content. No figure contains AI-generated imagery.
- **Author contributions:** S.A. conceived the study, designed the intervention catalogue and hybrid recommender, conducted the original analysis, developed the Engage prototype and wrote the first draft. T.J.A. supervised the project, designed

the extended evaluation, conducted the re-analysis and revised the manuscript. Both authors approved the final version.

References

[1] Hibbard, J.H., Greene, J.: What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Affairs **32**(2), 207–214 (2013) https://doi.org/10.1377/hlthaff.2012.1061

[2] Carman, K.L., Dardess, P., Maurer, M., Sofaer, S., Adams, K., Bechtel, C., Sweeney, J.: Patient and family engagement: a framework for understanding the elements and developing interventions and policies. Health Affairs **32**(2), 223–231 (2013) https://doi.org/10.1377/hlthaff.2012.1133

[3] West, L., Mitchell, D., Faulkner, S.D., Bauer, B., Brooke, N., Priest, E.: Digital health technologies: learnings and perspectives from a patient engagement stakeholder expectations matrix study. Journal of Medical Internet Research **27**, 81463 (2025) https://doi.org/10.2196/81463

[4] Zendehbad, S.A., Ghasemi, J., Samsami Khodadad, F.: Trustworthy AI in telehealth: navigating challenges, ethical considerations, and future opportunities for equitable healthcare delivery. Healthcare Technology Letters **12**(1), 70020 (2025) https://doi.org/10.1049/htl2.70020

[5] Guhan, P., Awasthi, N., McDonald, K., Bussell, K., Reeves, G., Manocha, D., Bera, A.: Developing a machine learning–based automated patient engagement estimator for telehealth: algorithm development and validation study. JMIR Formative Research **9**, 46390 (2025) https://doi.org/10.2196/46390

[6] Dullabh, P., Zott, C., Gauthreaux, N., Peterson, C., Aronoff, A., Monkhouse, K., Sittig, D.F.: Integrating generative AI into patient-centered clinical decision support: research and practice considerations. Journal of Medical Internet Research **28**, 81628 (2026) https://doi.org/10.2196/81628

[7] Monaco, F., *et al.*: An advanced artificial intelligence platform for a personalised treatment of eating disorders. Frontiers in Psychiatry **15**, 1414439 (2024) https://doi.org/10.3389/fpsyt.2024.1414439

[8] Hahne, J., Carpenter, B.D.: Patient and clinician engagement with generative artificial intelligence (GenAI): a scoping review. Patient Education and Counseling **145**, 109492 (2026) https://doi.org/10.1016/j.pec.2026.109492

[9] Wiesner, M., Pfeifer, D.: Health recommender systems: concepts, requirements, technical basics and challenges. International Journal of Environmental Research and Public Health **11**(3), 2580–2607 (2014) https://doi.org/10.3390/ijerph110302580

[10] Schäfer, H., Hors-Fraile, S., Karumur, R.P., Calero Valdez, A., Said, A., Torkamaan, H., Ulmer, T., Trattner, C.: Towards health (aware) recommender systems. In: Proceedings of the 2017 International Conference on Digital Health, pp. 157–161. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3079452.3079499

[11] Tran, T.N.T., Felfernig, A., Trattner, C., Holzinger, A.: Recommender systems in the healthcare domain: state-of-the-art and research issues. Journal of Intelligent Information Systems **57**, 171–201 (2021) https://doi.org/10.1007/s10844-020-00633-6

[12] Hors-Fraile, S., Rivera-Romero, O., Schneider, F., Fernandez-Luque, L., Luna-Perejon, F., Civit-Balcells, A., de Vries, H.: Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: a scoping review. International Journal of Medical Informatics **114**, 143–155 (2018) https://doi.org/10.1016/j.ijmedinf.2017.12.018

[13] De Croon, R., Van Houdt, L., Htun, N.N., Štiglic, G., Vanden Abeele, V., Verbert, K.: Health recommender systems: systematic review. Journal of Medical Internet Research **23**(6), 18035 (2021) https://doi.org/10.2196/18035

[14] Ferrari Dacrema, M., Cremonesi, P., Jannach, D.: Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 101–109. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3298689.3347058

[15] Sun, Z., Yu, D., Fang, H., Yang, J., Qu, X., Zhang, J., Geng, C.: Are we evaluating rigorously? Benchmarking recommendation for reproducible evaluation and fair comparison. In: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 23–32. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3383313.3412489

[16] Ji, Y., Sun, A., Zhang, J., Li, C.: A critical study on data leakage in recommender system offline evaluation. ACM Transactions on Information Systems **41**(3), 1–27 (2023) https://doi.org/10.1145/3569930

[17] Meng, Z., McCreadie, R., Macdonald, C., Ounis, I.: Exploring data splitting strategies for the evaluation of recommendation models. In: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 681–686. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3383313.3418479

[18] Krichene, W., Rendle, S.: On sampled metrics for item recommendation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1748–1757. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3403226

[19] Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill,

D., Pera, M.S.: All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 172–186 (2018)

[20] Abdollahpouri, H., Burke, R., Mobasher, B.: Controlling popularity bias in learning-to-rank recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 42–46. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3109859.3109912

[21] Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., McLachlan, S.: Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association **25**(3), 230–238 (2018) https://doi.org/10.1093/jamia/ocx079

[22] Steck, H.: Embarrassingly shallow autoencoders for sparse data. In: Proceedings of the World Wide Web Conference (WWW '19), pp. 3251–3257. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313710

[23] Glover, W., Renton, M., Minaye, H., Dabiri, O.: Assessing equitable development and implementation of artificial intelligence-enabled patient engagement technologies. Mayo Clinic Proceedings: Digital Health **3**(1), 100192 (2025) https://doi.org/10.1016/j.mcpdig.2024.100192

[24] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019) https://doi.org/10.1126/science.aax2342

[25] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., the Precise4Q consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**, 310 (2020) https://doi.org/10.1186/s12911-020-01332-6

[26] Dobson, R., Stowell, M., Whittaker, R.: Maintaining patient trust as artificial intelligence's role in healthcare grows. New Zealand Medical Journal **139**(1629), 94–101 (2026) https://doi.org/10.26635/6965.7122

[27] Lops, P., Gemmis, M., Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 73–105. Springer, Boston, MA (2011). https://doi.org/10.1007/978-0-387-85820-3_3

[28] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186. ACM, New York, NY, USA (1994). https://doi.org/10.1145/192844.192905

[29] Burke, R.: Hybrid recommender systems: survey and experiments. User Modeling and User-Adapted Interaction **12**(4), 331–370 (2002) https://doi.org/10.1023/A:1021240730564

[30] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002) https://doi.org/10.1145/582415.582418

[31] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems **22**(1), 5–53 (2004) https://doi.org/10.1145/963770.963772

[32] Gonzales, A., Guruswamy, G., Smith, S.R.: Synthetic data in health care: a narrative review. PLOS Digital Health **2**(1), 0000082 (2023) https://doi.org/10.1371/journal.pdig.0000082

[33] Chen, J., Chun, D., Patel, M., Chiang, E., James, J.: The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Medical Informatics and Decision Making **19**, 44 (2019) https://doi.org/10.1186/s12911-019-0793-0

[34] Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**(1), 75–105 (2004) https://doi.org/10.2307/25148625

[35] The MITRE Corporation: Synthea sample synthetic patient records, CSV format (April 2020 release). https://synthetichealth.github.io/synthea-sample-data/downloads/synthea_sample_data_csv_apr2020.zip. Accessed 25 October 2026. Apache License 2.0 (2020)

[36] Tintarev, N., Masthoff, J.: Designing and evaluating explanations for recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 479–510. Springer, Boston, MA (2011). https://doi.org/10.1007/978-0-387-85820-3_15

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1supplementarytables.pdf
- OnlineResource2codeandresults.zip