# Collaborative Filtering for the Imputation of Patient Reported Outcomes

Eric Ababio Anyimadu[1(✉)], Clifton David Fuller[2], Xinhua Zhang[3], G. Elisabeta Marai[3], and Guadalupe Canahuate[1]

[1] Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA
eric-anyimadu@uiowa.edu
[2] Department of Radiation Oncology, The University of Texas, MD. Anderson Cancer Center, Houston, TX, USA
[3] Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607, USA

**Abstract.** This study addresses the prevalent issue of missing data in patient-reported outcome datasets, particularly focusing on head and neck cancer patient symptom ratings sourced from the MD Anderson Symptom Inventory. Given that many data mining and machine learning algorithms necessitate complete datasets, the accurate imputation of missing data as an initial step becomes crucial. In this study we propose, for the first time, the use of collaborative filtering for imputing missing head and neck cancer patient symptom ratings. Two configurations of collaborative filtering, namely patient-based and symptom-based, leverage known ratings to infer the missing ones. Additionally, this study compares the performance of collaborative filtering with alternative imputation methods such as Multiple Imputation by Chained Equations, Nearest Neighbor Imputation, and Linear interpolation. Performance is compared using Root Mean Squared Error and Mean Absolute Error metrics. Findings demonstrate that collaborative filtering is a viable and comparatively superior approach for imputing missing patient symptom data.

**Keywords:** Head and Neck Cancer · Imputation · Collaborative Filtering

## 1 Introduction

Head and neck cancer (HNC) patients often experience disease-related symptoms and side effects during and after treatment which can affect their quality of life and survival [16]. Researchers and physicians are therefore increasingly placing significant attention on leveraging existing patient symptom data to personalize care for patients and improve patient outcomes [17]. Furthermore, the examination of patient symptom data has been recognized as having the capacity to yield fresh insights into clinical understanding to enhance diagnostic accuracy and optimize the effective allocation of healthcare resources [15].


C. Strauss et al. (Eds.): DEXA 2024, LNCS 14910, pp. 231–248, 2024.
https://doi.org/10.1007/978-3-031-68309-1_20

The MD. Anderson Symptom Inventory (MDASI) is a validated instrument to collect patient reported outcomes. The MDASI Head and Neck module (MDASI-HN) is a 28-symptom questionnaire relevant for head and neck cancer patients [20]. Patient responses are collected before, during, and after treatment and similar to other longitudinal datasets that rely on patient responses or feedback, the MDASI-HN data often contain missing values [1,24]. This imposes restrictions on the applicability of numerous statistical methods and machine learning approaches in analyzing these incomplete datasets, given that these techniques usually require complete datasets [5]. Moreover, discarding data from patients with missing responses in order to achieve complete datasets may introduce bias in parameter estimation. These patients could possess special characteristics that are not representative of the broader group, thus limiting the extent to which these analyses can be generalized [2,25]. To address this issue, missing values are commonly imputed as an initial step.

Several techniques exist for imputation, including Multiple Imputation by Chained Equations (MICE), K Nearest Neighbor (KNN) methods, and Linear Interpolation (LI) [7,11,24]. Despite their effectiveness in various scenarios, these methods may fail to capture intricate data relationships, particularly regarding patient sensitivity which are influenced by individual tolerance levels.

Collaborative filtering, a technique successfully employed in recommendation systems to leverage user preferences for personalized suggestions, offers a promising alternative [10]. We hence propose and evaluate the use of collaborative filtering to impute missing responses in the MDASI-HN, leveraging similarities in reported outcomes to enhance imputation accuracy.

Furthermore, we conduct experimental analyses comparing the performance of collaborative filtering against other established methodologies. Performance metrics used for evaluation include root mean square error and mean absolute error.

## 2 Related Work

We reviewed related work in two main categories: studies on the imputation of HNC symptom data and studies on collaborative filtering and its applications.

Imputation utilizes existing data and inherent associations to forecast specific or range-based approximations for missing values. Over the past few years, some studies have employed one imputation technique or the other in filling missing values in HNC symptom data. Some of the widely used imputation techniques are MICE, KNN and LI. MICE iteratively fills missing values in a dataset, creating a complete set of data in each cycle, improving with each iteration until an ultimate dataset is achieved [11,13]. Conversely, KNN leverages intrinsic patient similarities to infer missing outcomes, while LI estimates values assuming a linear relationship [24].

Relevant studies in this field include one focused on the impact of radiation-induced toxicities on the quality of life for patients treated for HNC, which utilized MICE to complete both physician-rated toxicities and patient-rated symptoms post-radiotherapy [11]. Another study on using a Long Short-Term Memory

(LSTM) neural network to predict late-stage symptom severity demonstrated the effectiveness of imputation techniques, including LI and MICE, for addressing missing data in MDASI-HN patient-reported outcomes [24]. Additionally, a study on predicting clinical outcomes of radiotherapy in HNC patients employed statistical, MICE, and KNN imputation methods, highlighting the superior performance of MICE compared to the other techniques [7].

While these and other studies have employed various techniques to fill missing values, they primarily used these methods as pre-processing steps without focusing on comprehensive evaluations of the imputation techniques.

Collaborative filtering (CF) methods are widely used in recommendation systems such as GroupLens, Amazon.com, Netflix, Google News, and Facebook and excel in predicting user preferences based on collected ratings [19]. CF methods have been proposed for data imputation as well. The auto-adaptive CF imputation method, which leverages both item and user ratings to predict missing values and validated using the MovieLens dataset, was shown to outperform traditional imputation techniques [14]. Similarly, CF method based on rough-set theory was applied for imputing missing values in microarray gene expression data [23]. This CF based method outperformed KNN method over changing rates of missing values.

These studies showed the viability of CF in imputing missing values in a wide variety of fields. Nonetheless, to the best of our knowledge, CF has not been applied for the imputation of MDASI-HN patient-reported outcomes data before. Our objective is to demonstrate the effectiveness of collaborative filtering compared to established methods in this context. We employ traditional CF methods as a foundation, paving the way for future research on this topic using MDASI-HN data.

## 3 Methodology

In this section, we begin by describing the data. Next, we introduce the collaborative filtering technique and explain how we used it to fill missing patient-reported outcomes.

### 3.1 MDASI-HN Data

The MDASI-HN 28 questionnaire items are categorized as follows: 13 core MDASI items that rate general cancer symptoms, 9 HNC-specific items that rate symptoms associated with HNC and 6 interference items that assess how severely symptoms interfere with daily activities [20].

Each patient self-reports the 28 symptoms on a 0–10 scale with 0 indicating "not present" and 10 indicating "as bad as you can imagine". Patients are asked to rate each item according to its worst severity during the previous 24 h [21].

All HNC patients in the cohort underwent standard of care treatment (radiotherapy with or without chemotherapy) with curative intent. The HNC patients completed the MDASI-HN questionnaires at the following stages: baseline ratings before the start of treatment, weekly evaluations spanning 7 weeks throughout the treatment course, and additional assessments after the $6^{th}$ week as well as at the $6^{th}$, $12^{th}$, and $18^{th}$ months after completion of treatment. The MDASI questionnaires can be abstracted as a two-dimensional user-item matrix where rows correspond to patients and columns correspond to symptoms.

We denote as $R_{p,i}$ the rating for patient $p$ and symptom $i$. We distinguish between different time points, denoting as $R^t_{p,i}$, the rating provided by patient $p$ for symptom $i$ at time point $t$. For patients with missing ratings, $R^t_{p,i} = NA$.

In addition to symptom data, the dataset also includes clinical information of each patient such as biographic information (age, sex, change in height and weight during treatment), disease specifics (site of tumor, new disease after primary and TNM stage) and treatment information (prior treatment at enrolment, induction or concurrent chemotherapy, neck dissection and surgery status).

### 3.2 Collaborative Filtering (CF) for MDASI-HN

CF methods leverage the similarity between known preferences of the users without requiring the use of other external information to predict unknown preferences [10,22]. There are two variations of the CF techniques: the user-based which leverages similarity between users and the item-based method which exploits the similarity between items [22].

Let's consider an example using the user-based CF approach for book recommendations. The users provide book ratings to indicate their book preferences (e.g. likes and dislikes). Given the current preferences of a user $p$ and the preferences of all other users, we are seeking to predict whether user $p$ would like book $i$. The first step is to identify users that have rated book $i$ and select the top $k$ users ranked by the similarity of their preferences to the preferences of $p$. The average rating from the $k$ users is used to predict the rating user $p$ would give to book $i$. For the item-based approach, all the existing ratings for book $i$ are compared against the ratings for all other books user $p$ has rated and the top $k$ most similar ones are used to predict the rating for book $i$ for user $p$.

We adapt the user-based and item-based CF approaches to derive the CF Patient-based and CF Symptom-based methods to predict missing symptom ratings as explained below.

**CF Patient-Based (CF-PAT)** approach predicts missing ratings using known ratings from other patients who are most similar to the given patient. The procedure to impute a missing rating for symptom $i$ at time-point $t$ by patient $p$ represented as $R^t_{p,i}$ using CF-PAT imputation is as follows:

- Find all patients $Q$ who have known ratings for symptom $i$ at time-point $t$.
- Determine the similarity, $\text{sim}(p, q)$, between patient $p$ and each $q \in Q$ using their common known ratings.
- Select the top $k$ of these $q$, i.e. $q[1], ...., q[k]$ patients that are most similar to $p$.
- Calculate the missing rating as the average of the ratings of symptom $i$ at time-point $t$ by $k$ weighted by their similarity measure as shown in Eq. 1:

$$R^t_{p,i} = \frac{\sum_k R^t_{q[k],i} \ * \ Sim(p, q[k])}{\sum_k Sim(p, q[k])} \tag{1}$$

where $Sim(p, q)$ is a patient similarity derived using a similarity metric (see Table 1) over the common ratings between patients $p$ and $q$.

**CF Symptom-Based (CF-SYM)** on the other hand predicts missing ratings using known ratings from other selected symptoms or time-points rated by the same patient. This selection is guided by the inter-symptom relationships identified across all patients. The process to impute missing rating for symptom $i$ at time-point $t$ by patient $p$ denoted as $R^t_{p,i}$ is derived as follows:

- Find all symptoms $J$ where ratings for patient $p$ are known.
- Using a similarity metric, determine the similarity measure between symptom $i$ and all $j \in J$ symptoms using their common known ratings among all patients.
- Select the top $k$ of $j$, i.e. $j[1], ...., j[k]$ that are most similar to $i$ using their similarity measures.
- Impute the missing rating as the average ratings of the $k$ symptoms rated by $p$ weighted their similarity measure to symptom $i$ as shown in Eq. 2:

$$R^t_{p,i} = \frac{\sum_k R^*_{p,j[k]} \ * Sim(i, j[k])}{\sum_k Sim(i, j[k])} \tag{2}$$

where $Sim(i, j)$ is a symptom similarity derived using a similarity metric (see Table 1) over the common ratings between symptoms $i$ and $j$. The notation $R^*_{p,j}$ is used to indicate that each symptom time point is considered independently when computing the similarity between symptoms and the most similar time points are used for rating imputation.

In both the CF-PAT and CF-SYM configurations, the top 5 most similar neighboring patients or symptoms are selected (i.e. $k = 5$). The process described above is repeated until all missing ratings are filled.

**Table 1. CF Similarity Metrics.** For Patient-based similarity, N represents the count of shared ratings between patients $p$ and $q$. The symbols $\hat{R}_p$ and $\hat{R}_q$ used in PCC denote the average ratings of the shared ratings between patients $p$ and $q$, respectively. For Symptom-based similarity, N signifies the number of patients who have rated both symptoms $i$ and $j$, and $\hat{R}_i$ and $\hat{R}_j$ represent the averages of the common ratings between symptoms $i$ and $j$, respectively.

| | Patient-based | Symptom-based |
|---|---|---|
| Euclidean similarity | CF-PAT-EUC | CF-SYM-EUC |
| | $\text{Sim}(p,q) = 1 - \sqrt{\sum_{i=1}^{N}(R_{p,i} - R_{q,i})^2}$ | $\text{Sim}(i,j) = 1 - \sqrt{\sum_{p=1}^{N}(R_{p,i} - R_{p,j})^2}$ |
| Cosine similarity | CF-PAT-COS | CF-SYM-COS |
| | $\text{Sim}(p,q) = \frac{\sum_{i=1}^{N} R_{p,i} \cdot R_{q,i}}{\sqrt{\sum_{i=1}^{N}(R_{p,i})^2 \cdot (R_{q,i})^2}}$ | $\text{Sim}(i,j) = \frac{\sum_{p=1}^{N} R_{p,i} \cdot R_{p,j}}{\sqrt{\sum_{p=1}^{N}(R_{p,i})^2 \cdot (R_{p,j})^2}}$ |
| Pearson correlation coefficient | CF-PAT-PCC | CF-SYM-PCC |
| | $\text{Sim}(p,q) = \frac{\sum_{i=1}^{N}(R_{p,i} - \hat{R}_p) \cdot (R_{q,i} - \hat{R}_q)}{\sqrt{\sum_{i=1}^{N}(R_{p,i} - \hat{R}_p)^2 \cdot (R_{q,i} - \hat{R}_q)^2}}$ | $\text{Sim}(i,j) = \frac{\sum_{p=1}^{N}(R_{p,i} - \hat{R}_i) \cdot (R_{p,j} - \hat{R}_j)}{\sqrt{\sum_{p=1}^{N}(R_{p,i} - \hat{R}_i)^2 \cdot (R_{p,j} - \hat{R}_j)^2}}$ |

**Similarity Metrics:** We experimented with three commonly used similarity metrics in both the CF-PAT and CF-SYM techniques [9]. These similarity metrics were Euclidean similarity (EUC), the vector-based Cosine similarity (COS) and the correlation-based Pearson Correlation Coefficient similarity (PCC) [9,22].

EUC is a linear metric and has gained widespread applicability due to its simplicity and effectiveness, particularly in the analysis of non-sparse numerical data [9]. Meanwhile, COS approach treats sets of ratings as vectors, calculating the cosine angle between them. This method carries the advantage of naturally providing a normalized distance measure. PCC also measures the linear relationship between two sets of ratings, expressed as the ratio of their covariance to the standard deviation [9,22].

Each of these similarity measures contributed uniquely to the analyses, catering to different aspects of similarity evaluation in the dataset. Table 1 provides the different equations used to determine the various measures of similarity using EUC, COS or PCC in the CF-PAT or CF-SYM configurations.

Note that to ensure consistency, all similarity values were normalized to range from 0, signifying no similarity, to 1, representing the highest degree of similarity.

## 4 Evaluation

To assess the performance of the imputation techniques, in addition to the original missing values, we randomly masked some known values to serve as our ground truth data per symptom. We assumed that patients provided at least one rating for each symptom throughout the monitoring period and hence during the masking process, we ensured that every patient retained at least one known rating for each symptom.

We evaluated both the patient-based (CF-PAT) and symptom-based (CF-SYM) versions of the collaborative filtering (CF) methods, employing the three distinct similarity measures: Euclidean distance (EUC), Cosine similarity (COS), and Pearson correlation coefficient (PCC). Furthermore, we explored more adaptations of CF-PAT, considering the different treatment stages (baseline, during treatment, and post-treatment) independently, which we termed Per Treatment Stage (PTS).

Also, a prerequisite for calculating the similarity was to have an arbitrary minimum of 10 common ratings between patients or symptoms to ensure reliability of the measurements.

We compared the performance of the nine CF-based methods against three established methods: MICE, KNN imputation, and LI.

We applied the MICE technique with two different configurations: MICE with clinical data (MICE-w-Clinical) and MICE with only ratings (MICE-Ratings). The KNN method computed similarity between patients using Euclidean similarity over the available clinical data. Additionally, the LI method filled missing values for each patient and symptom independently, leveraging known patient ratings for a given symptom at various time points.

### 4.1 Evaluation Metrics:

As is commonly used in evaluating machine learning models, we assessed imputation performance using root mean squared error (RMSE) and mean absolute error (MAE) measurements [8].

MAE is a linear error measurement, implying that all individual deviations are assigned equal importance in determining the overall result making it a more natural measure of average error. On the other hand, RMSE calculates the average magnitude of squared errors and consequently assigns relatively higher weight to larger errors making it comparatively more sensitivity [8].

RMSE and MAE over $T$ imputation points are derived as shown in Eqs. 3 and 4 respectively:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(P_t - A_t)^2} \tag{3}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\|P_t - A_t\| \tag{4}$$

where $P$ and $A$ are the sets of imputed and actual/ground truth data respectively. Smaller values of RMSE/MAE indicate a better performance.

# 5 Experimental Results

## 5.1 Experimental Setup

To inject missing values in the data, random masking was performed to remove an average of 2%-3% of the original values. All missing values were then imputed, and using the ground truth from the masked values, the imputation methods were evaluated using RMSE and MAE.

We repeated the experiments five times, each time randomly generating masked versions of the dataset and reported the average evaluation metric scores for each method.

We performed all the analyses using python 3.11 version. Python scikit learn, numpy and pandas were used for data pre-processing and implementation of the imputation techniques. The experiments were conducted on a MacBook Pro.

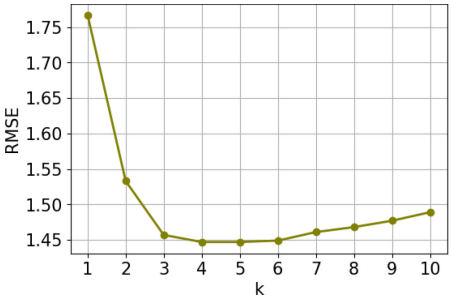
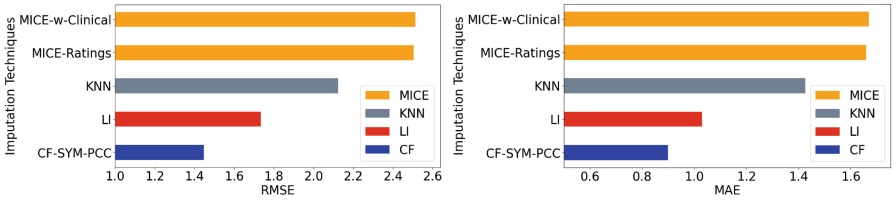



**Fig. 1.** RMSE of CF-SYM-PCC imputation over changing number of selected neighbors ($k$).




(a) RMSE (b) MAE

**Fig. 2. Comparison of the imputation techniques.** Over the masked or ground truth dataset, CF-SYM-PCC was the best imputation method using either (a) RMSE or (b) MAE values.

## 5.2 Data Statistics

The data for the analyses encompassed a cohort of 821 patients, and Table 2 provides a breakdown of the distribution and frequency of missing symptom

ratings among these patients, stratified based on their clinical data. To ensure consistency and eliminate discrepancies arising from diverse measurement scales all categorical clinical records were transformed into binary representations using one-hot encoding, while numerical values were normalized to a range of 0 to 1 before use in the analyses [18].

**Table 2.** Cohort Distribution and Missing Symptom Rate Stratified By Clinical Data.

| Features | Categories | Distribution of Patients (%) | Rate of Missing Ratings (%) |
|---|---|---|---|
| **Biographical Data** | | | |
| Age | <60 years | 47.25 | 20.50 |
| | ≥ 60 years | 52.75 | 23.00 |
| Sex | Female | 11.32 | 5.27 |
| | Male | 88.68 | 38.10 |
| Height change during treatment | Increase | 2.10 | 0.81 |
| | Decrease | 1.06 | 0.59 |
| | No change | 96.84 | 42.45 |
| Weight change during treatment | Increase | 4.50 | 1.94 |
| | Decrease | 32.28 | 11.95 |
| | No change | 63.22 | 29.94 |
| **Disease Data** | | | |
| Site of Tumor | Base of tongue | 44.62 | 18.06 |
| | Tonsil | 44.00 | 19.44 |
| | Others | 4.37 | 46.41 |
| | Not specified | 7.01 | 2.75 |
| New disease after primary | Yes | 5.11 | 1.06 |
| | No | 94.89 | 15.39 |
| T-stage | t0 | 7.08 | 2.56 |
| | t1 | 29.33 | 11.50 |
| | t2 | 37.42 | 15.55 |
| | t3 | 13.65 | 5.84 |
| | t4 | 11.88 | 5.63 |
| | tx | 0.64 | 0.26 |
| N-stage | n0 | 12.77 | 6.51 |
| | n1 | 34.77 | 14.31 |
| | n2,a,b,c | 49.94 | 40.43 |
| | n3,a | 2.15 | 50.95 |
| | nx | 0.37 | 0.06 |
| M-stage | 1 | 5.04 | 2.11 |
| | 2 | 94.96 | 32.79 |

(*continued*)

**Table 2.** (*continued*)

| Features | Categories | Distribution of Patients (%) | Rate of Missing Ratings (%) |
|---|---|---|---|
| **Treatment Data** | | | |
| Status at Enrollment | Previously Treated | 5.16 | 2.20 |
| | Previously Untreated | 94.84 | 41.28 |
| Induction Chemotherapy | Yes | 23.42 | 9.36 |
| | No | 76.58 | 26.78 |
| Concurrent Chemotherapy | Yes | 59.21 | 23.65 |
| | No | 40.79 | 17.90 |
| Neck Dissection | Yes | 75.27 | 9.46 |
| | No | 24.73 | 24.18 |
| Surgery at Primary Site | Yes | 80.22 | 8.28 |
| | No | 19.78 | 25.15 |

The rate of missing symptoms originally in the data and average rate over five random masks according to the treatment stages are as follows: baseline (original: 19.66%, after masking: 21.71%), during treatment (original: 50.57%, after masking: 53.05%) and post-treatment (original: 42.89%, after masking: 45.38%).

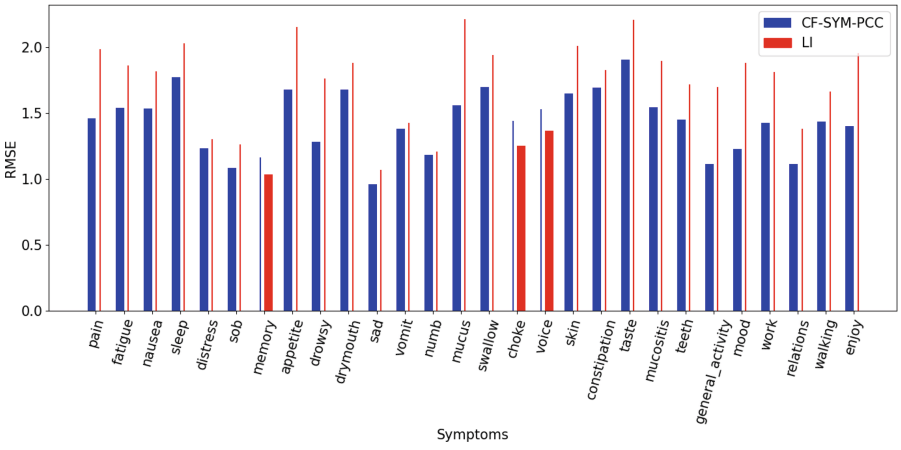



**Fig. 3. RMSE comparison between CF-SYM-PCC and LI methods per symptom.** The best (smaller) RMSE for each symptom is represented by a thicker bar.

### 5.3 CF Techniques Comparison

Table 3 shows the RMSE and MAE results for the evaluation of the nine variations of CF techniques, sorted by MAE from best to worse. In general, the CF-SYM techniques outperformed the CF-PAT methods. This superior performance of CF-SYM can be attributed to several factors. Firstly, the association between symptoms tends to be more established than that between patients as symptoms often exhibit clearer patterns of co-occurrence [6,26]. This is further supported by research which indicates that symptom-based models can very effectively capture underlying disease and treatment responses [4]. Additionally, each symptom at a given time point has more ratings from individual patients compared to the number of ratings provided by each patient for the fewer number of symptoms. Consequently, CF-SYM leverages a larger number of ratings for determining similarity relative to CF-PAT, resulting in a more reliable and robust selection of high-quality neighbors or collaborators and hence better accuracy of the CF-SYM imputations.

In terms of the similarity metrics, PCC emerged as the optimal for CF-SYM. As compared to the other metrics, PCC in determining similarity normalizes all ratings by subtracting the mean of common ratings between each pair of symptoms. This mean normalization scales all ratings used in computing the PCC similarity uniformly therefore making the levels of ratings comparable regardless of the actual numeric values. Mean scaling therefore reduces the variations between ratings and ensures similar patterns in symptom ratings are identified for imputation. As a result, PCC is more effective at identifying nuanced correlations that might be overlooked by EUC and COS, leading to more precise and reliable similarity assessments in CF-SYM.

Overall, the best CF method for missing value imputation was CF-SYM-PCC under both RMSE and MAE metrics. Therefore, for the rest of this section, we focus on the performance of CF-SYM-PCC and its comparison with other methods.

**Table 3.** CF Techniques Comparison Results

| CF Techniques | RMSE | MAE |
|---|---|---|
| **CF-SYM-PCC** | **1.447** | **0.898** |
| CF-SYM-EUC | 1.738 | 0.998 |
| CF-SYM-COS | 1.739 | 1.023 |
| CF-PAT-EUC-PTS | 1.737 | 1.048 |
| CF-PAT-EUC | 1.843 | 1.107 |
| CF-PAT-PCC | 1.797 | 1.127 |
| CF-PAT-PCC-PTS | 1.813 | 1.171 |
| CF-PAT-COS-PTS | 1.842 | 1.239 |
| CF-PAT-COS | 1.842 | 1.239 |

### 5.4 Effect of $k$ on CF-SYM-PCC Imputation

Figure 1 illustrates the effect of varying the number of selected neighbors, represented by $k$, on the performance of the CF-SYM-PCC technique. Notably, the optimal performance is seen for $k$ being 4 or 5 (RMSE: 1.447), while the least desirable performance was observed at $k = 1$ (RMSE: 1.767). The pattern follows the observed behavior of KNN approaches in other works. When $k$ values are exceedingly small, collaboration effectiveness may be constrained. Conversely, larger $k$ values beyond a certain threshold can potentially distort the original data variations and dilute the influence of genuine collaborators [3].

Therefore, for the rest of our experiments, we use $k = 5$ as it provided the optimal selection of correlated symptoms for imputing a missing symptom for this data.

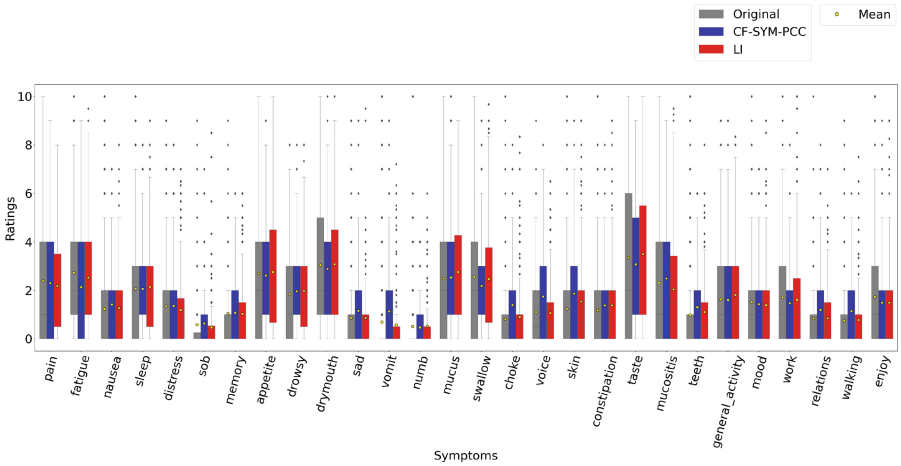



**Fig. 4. Box plots of the randomly selected known ratings that were masked for evaluation per symptom.** Original represents the pre-substituted ratings while the CF-SYM-PCC and LI represent the predicted or imputed values using the respective techniques.

### 5.5 Comparing CF-SYM-PCC Against Other Methods

Figure 2 shows the RMSE and MAE comparison between the proposed method (CF-SYM-PCC) and other popular imputation techniques. The results are ordered by descending RMSE and MAE values on the vertical axis, so the best performing method is shown at the bottom.

As can be seen, the CF-SYM-PCC technique was the best performing method with the lowest error rates in both RMSE and MAE metrics (RMSE: 1.447, MAE: 0.898). These results support that leveraging symptom-based collaborative filtering with the Pearson correlation coefficient as the similarity measure is an effective method for PRO data imputation.

The simple Linear Interpolation (LI) method shows the second best comparative performance (RMSE: 1.734, MAE: 1.029) and it is only out-performed by the CF-SYM-PCC method. The good performance of the LI method can be attributed to the fact that symptom ratings are temporally correlated.

Both MICE and KNN imputation had the worst performance with the highest errors (MICE-w-Clinical [RMSE: 2.513, MAE:1.67]), (MICE-Ratings [RMSE: 2.505, MAE: 1.659]), and KNN Imputation (RMSE: 2.123, MAE:1.425). This indicates that clinical features are not very effective in predicting symptom ratings for these patients. There are local correlations between symptoms (e.g. symptom clusters) that using clinical features are not exploited.

The homogeneous nature of the cohort, which received similar treatment regimens for HNC and hence experienced similar symptoms, likely explains the success of the CF method and the relatively limited performance of the clinical data-based methods such as MICE-w-Clinical and KNN. The performance of the techniques may possibly vary under diverse scenarios involving heterogeneous cohorts.

Furthermore, given that symptom ratings often exhibit linearity, it is not surprising that LI and CF-SYM-PCC, which rely heavily on a patient's own ratings, demonstrated relatively superior performance compared to other methodologies.

In the next section, we proceed to compare the performance of the top two methods: LI and CF-SYM-PCC on a per symptom basis.

### 5.6 Comparing CF-SYM-PCC and LI Techniques Per Symptom

Figure 3 shows the RMSE for the CF-SYM-PCC and LI methods. Each column represents a symptom with two bars. The thicker bar corresponds to the best performing method for that symptom while the thin bar is the RMSE of the other method included for comparison. As can be seen, CF-SYM-PCC had better performance across all the symptoms except for memory, choke and voice. Taste had the highest RMSE (CF-SYM-PCC: 1.909, LI: 2.209) overall, while sadness had the lowest RMSE (CF-SYM-PCC: 0.960, LI: 1.072). Table 4 reports, in addition to the overall RMSE for each symptom and both methods, the per treatment stage RMSE for baseline, during treatment, and after treatment.

Figure 4 demonstrates the spread of both the originally masked and corresponding imputed data, predicted by the CF-SYM-PCC and LI methods. The predictions by both techniques were within the interval of symptom rating values. Also, while both techniques introduced small mean shifts, the distribution of the imputed data is within acceptable ranges from the ground truth, as evidenced by the figure.

**Table 4.** RMSE per symptom over treatment stages.

| Symptoms | Overall | | Baseline | | During Treatment | | Post Treatment | |
|---|---|---|---|---|---|---|---|---|
| | CF-SYM-PCC | LI | CF-SYM-PCC | LI | CF-SYM-PCC | LI | CF-SYM-PCC | LI |
| Pain | 1.463 | 1.988 | 1.594 | 2.38 | 1.450 | 2.186 | 1.444 | 1.399 |
| Fatigue | 1.542 | 1.861 | 1.966 | 3.235 | 1.446 | 1.691 | 1.604 | 1.767 |
| Nausea | 1.534 | 1.817 | 1.106 | 3.037 | 1.82 | 2.043 | 0.735 | 0.905 |
| Sleep | 1.772 | 2.033 | 1.558 | 2.554 | 1.821 | 1.821 | 1.728 | 2.259 |
| Distress | 1.231 | 1.303 | 1.969 | 1.791 | 1.127 | 1.300 | 1.153 | 1.122 |
| SOB | 1.084 | 1.261 | 1.581 | 2.372 | 0.823 | 0.808 | 1.365 | 1.607 |
| Memory | 1.162 | 1.034 | 1.394 | 0.745 | 1.111 | 0.938 | 1.195 | 1.257 |
| Appetite | 1.681 | 2.154 | 1.398 | 2.576 | 1.705 | 2.116 | 1.706 | 2.100 |
| Drowsy | 1.282 | 1.766 | 1.363 | 2.299 | 1.437 | 1.752 | 0.893 | 1.612 |
| Drymouth | 1.682 | 1.881 | 1.414 | 2.930 | 1.64 | 1.661 | 1.837 | 1.869 |
| Sad | 0.96 | 1.072 | 1.301 | 2.019 | 0.883 | 0.903 | 0.985 | 0.907 |
| Vomit | 1.382 | 1.425 | 0.707 | 1.087 | 1.549 | 1.672 | 1.184 | 0.938 |
| Numb | 1.182 | 1.208 | 0.845 | 0.787 | 0.862 | 0.859 | 1.559 | 1.618 |
| Mucus | 1.56 | 2.214 | 0.816 | 1.751 | 1.430 | 1.887 | 1.818 | 2.681 |
| Swallow | 1.701 | 1.943 | 2.356 | 2.145 | 1.676 | 1.980 | 1.530 | 1.820 |
| Choke | 1.44 | 1.253 | 1.026 | 0.229 | 1.570 | 1.404 | 1.294 | 1.128 |
| Voice | 1.533 | 1.368 | 1.309 | 1.488 | 1.692 | 1.398 | 1.237 | 1.290 |
| Skin | 1.651 | 2.010 | 1.072 | 0.837 | 1.900 | 2.020 | 1.191 | 2.193 |
| Constipation | 1.695 | 1.827 | 1.399 | 2.331 | 1.751 | 1.835 | 1.679 | 1.661 |
| Taste | 1.909 | 2.209 | 1.118 | 3.102 | 1.957 | 2.220 | 1.946 | 1.923 |
| Mucositis | 1.545 | 1.897 | 1.472 | 0.782 | 1.492 | 1.938 | 1.643 | 1.983 |
| Teeth | 1.453 | 1.718 | 1.103 | 1.504 | 1.560 | 1.837 | 1.350 | 1.566 |
| General activity | 1.116 | 1.701 | 1.078 | 1.626 | 1.277 | 1.864 | 0.630 | 1.288 |
| Mood | 1.226 | 1.882 | 1.432 | 2.000 | 1.185 | 1.928 | 1.254 | 1.744 |
| Work | 1.426 | 1.814 | 1.155 | 1.125 | 1.697 | 1.956 | 0.833 | 1.665 |
| Relations | 1.114 | 1.381 | 0.938 | 2.28 | 1.305 | 1.250 | 0.761 | 1.266 |
| Walking | 1.435 | 1.665 | 1.323 | 0.707 | 1.553 | 1.693 | 1.21 | 1.743 |
| Enjoy | 1.4 | 1.957 | 1.704 | 3.338 | 1.433 | 1.885 | 1.244 | 1.606 |

### 5.7 PCC Correlation Symptom Clusters

Figures 5 is a heat map showing the normalized Pearson correlation coefficients and clustering patterns among symptoms in the CF-PCC-SYM post-imputed dataset. The inter-symptom correlations shown in these figures are computed by averaging the correlations across corresponding time-points for each symptom

pair. These correlation values represent the strength of the relationship between each pair of symptoms. The dendrograms were generated using agglomerative hierarchical clustering and the inter symptom correlation as distance. Initially, each symptom is placed into its own cluster and the highest correlated symptoms are merged first. These clusters are subsequently expanded by averaging the distance between members and other candidate symptoms. The clustering proceeds until all symptoms are in the same cluster.

As can be seen in the figure, there are some strong clusters. These clusters are also evident in the pre-imputed data. These clusters are intuitive and have been identified by prior studies [6,12,15,21,26]. For example, all interference symptoms {general_activity, walking, work, relations, mood, and enjoy} are clustered together along {distress and sad}. The {appetite, sleep, fatigue, drowsy} is an intuitive cluster, as well as {nausea, vomit} which was strengthened after imputation. The {dry mouth and taste} cluster has also been reported previously together with the {taste, choke, voice, mucus, swallow, pain, and mucositis} cluster [12,26].

These results show that the CF-SYM-PCC imputed MDASI-HN dataset preserves certain well-established inter-symptom associations or clusters.

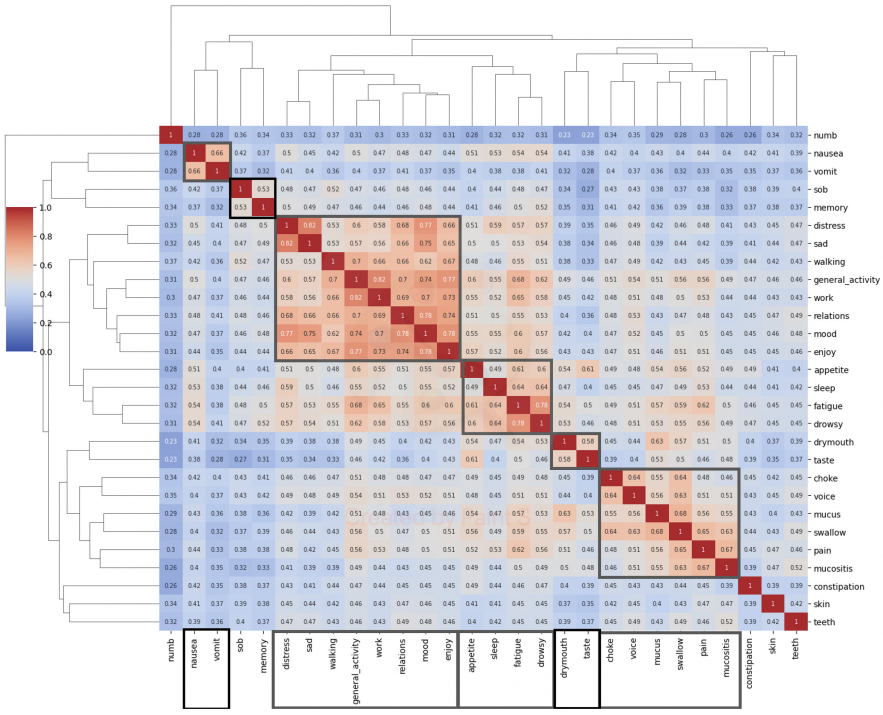



**Fig. 5. Average Pearson Correlation Coefficient between every pair of symptoms (after CF imputation).** The boxes around the diagonals indicate symptom clusters identified using agglomerative hierarchical clustering. These are consistent with prior literature in symptom cluster analysis and clinically relevant.

## 6 Conclusion

Collaborative filtering is an effective approach used in recommendation systems to leverage user preferences and recommend items that the user is likely to buy or consume. In this work, we have demonstrated that collaborative filtering can be applied to patient reported outcome data to provide a new and competitive approach for imputing patient data. In our experiments using HNC MDASI-HN data, the best performing configuration of the CF technique was the one denoted as CF-SYM-PCC which use item-based CF and the Pearson Correlation Coefficient for symptom similarity. This CF technique had the best overall (smallest) RMSE and MAE values among all the imputation methods considered, including MICE, KNN imputation, and linear interpolation. Linear interpolation was the second best performing method, and when compared on a per symptom basis, CF-SYM-PCC outperformed Linear Interpolation for 25 out of the 28 symptoms. We partly attribute the excellent performance of the CF method to the homogeneous nature of the cohort, which are all oropharyngeal cancer patients that received similar treatment regimens and hence expected to experience similar symptoms. Evaluating the performance of the CF techniques under diverse scenarios involving heterogeneous cohorts is left as subject for future work.

**Acknowledgement.** This work was supported directly or in part by funding/resources from the National Institutes of Health (NIH) National Cancer Institute (R01CA258827); received infrastructure support from the MD Anderson Cancer Center Support Grant Head and Neck Cancer Program and Image-Driven Biologically-informed Therapy (IDBT) Program, with programmatic support from the University of Texas MD Anderson Cancer Center Charles and Daneen Stiefel Center for Head and Neck Cancer Oropharyngeal Cancer Research Program; and the MD Anderson Image-guided Cancer Therapy Program.

We further acknowledge Amy C. Moreno, PhD, and Katherine A, Hutcheson, PhD of the STIEFEL program.

## References

1. Ayilara, O.F., Zhang, L., Sajobi, T.T., Sawatzky, R., Bohm, E., Lix, L.M.: Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual. Life Outcomes **17**, 1–9 (2019)
2. Bell, M.L., Fairclough, D.L.: Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat. Methods Med. Res. **23**(5), 440–459 (2014)
3. Beretta, L., Santaniello, A.: Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Mak. **16**, 197–208 (2016)
4. Bhagwat, N., Viviano, J.D., Voineskos, A.N., Chakravarty, M.M., Initiative, A.D.N., et al.: Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. PLoS Comput. Biol. **14**(9), e1006376 (2018)
5. Caiafa, C.F., Sun, Z., Tanaka, T., Marti-Puig, P., Solé-Casals, J.: Machine learning methods with noisy, incomplete or small datasets (2021)

6. Chiang, S., Ho, K., Wang, S.Y., Lin, C.: Change in symptom clusters in head and neck cancer patients undergoing postoperative radiotherapy: a longitudinal study. Eur. J. Oncol. Nurs. **35**, 62–66 (2018)
7. Gangil, T., Shahabuddin, A.B., Dinesh Rao, B., Palanisamy, K., Chakrabarti, B., Sharan, K.: Predicting clinical outcomes of radiotherapy for head and neck squamous cell carcinoma patients using machine learning algorithms. J. Big Data **9**(1), 25 (2022)
8. Hodson, T.O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. **15**(14), 5481–5487 (2022)
9. Jain, G., Mahara, T., Tripathi, K.N.: A survey of similarity measures for collaborative filtering-based recommender system. In: Pant, M., Sharma, T., Verma, O., Singla, R., Sikander, A. (eds.) Soft Computing: Theories and Applications: Proceedings of SoCTA 2018, pp. 343–352. Springer, Heidelberg (2020). https://doi.org/10.1007/978-981-15-0751-9_32
10. Koren, Y., Rendle, S., Bell, R.: Advances in collaborative filtering. In: Recommender Systems Handbook, pp. 91–142 (2021)
11. van der Laan, H.P., Van den Bosch, L., Schuit, E., Steenbakkers, R.J., van der Schaaf, A., Langendijk, J.A.: Impact of radiation-induced toxicities on quality of life of patients treated for head and neck cancer. Radiother. Oncol. **160**, 47–53 (2021)
12. Li, Y., et al.: Symptom clusters in head and neck cancer patients with endotracheal tube: Which symptom clusters are independently associated with health-related quality of life? Eur. J. Oncol. Nurs. **48**, 101819 (2020)
13. Luo, Y., Szolovits, P., Dighe, A.S., Baron, J.M.: 3D-mice: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. J. Am. Med. Inform. Assoc. **25**(6), 645–653 (2018)
14. , Ma, H., King, I., Lyu, M.R.: Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–46 (2007)
15. Mathew, A., et al.: Symptom clusters in head and neck cancer: a systematic review and conceptual model. In: Seminars in Oncology Nursing, vol. 37, p. 151215. Elsevier (2021)
16. Morton, R.P., Izzard, M.E.: Quality-of-life outcomes in head and neck cancer patients. World J. Surg. **27**, 884–889 (2003)
17. Noel, C.W., et al.: Enhancing outpatient symptom management in patients with head and neck cancer: a qualitative analysis. JAMA Otolaryngol.-Head Neck Surg. **148**(4), 333–341 (2022)
18. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. Int. J. Comput. Appl. **175**(4), 7–9 (2017)
19. Raghuwanshi, S.K., Pateriya, R.: Collaborative filtering techniques in recommendation systems. Data Eng. Appl. **1**, 11–21 (2019)
20. Rosenthal, D.I., et al.: Measuring head and neck cancer symptom burden: the development and validation of the md Anderson symptom inventory, head and neck module. Head Neck: J. Sci. Special. Head Neck **29**(10), 923–931 (2007)
21. Shi, Q., Mendoza, T.R., Gunn, G.B., Wang, X.S., Rosenthal, D.I., Cleeland, C.S.: Using group-based trajectory modeling to examine heterogeneity of symptom burden in patients with head and neck cancer undergoing aggressive non-surgical therapy. Qual. Life Res. **22**, 2331–2339 (2013)
22. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009** (2009)

23. Wang, B.W., Tseng, V.S., et al.: Improving missing-value estimation in microarray data with collaborative filtering based on rough-set theory. Int. J. Innov. Comput. Inf. Control **8**(3), 2157–2172 (2012)
24. Wang, Y., et al.: Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes. In: Proceedings of the 25th International Database Engineering & Applications Symposium, pp. 273–279 (2021)
25. Weber, G.M., et al.: Biases introduced by filtering electronic health records for patients with "complete data". J. Am. Med. Inform. Assoc. **24**(6), 1134–1141 (2017)
26. Xiao, C., et al.: Symptom clusters in patients with head and neck cancer receiving concurrent chemoradiotherapy. Oral Oncol. **49**(4), 360–366 (2013)