Original Papers

Usability and Acceptability of a Palliative Care Mobile Intervention for Older Adults With Heart Failure and Caregivers: Observational Study (e35592)
Jennifer Villalobos, Sheana Bull, Jennifer Portz. ................................................................. 2

Evolving Hybrid Partial Genetic Algorithm Classification Model for Cost-effective Frailty Screening: Investigative Study (e38464)
John Oates, Niusha Shaflabady, Rachel Ambagsheer, Justin Beilby, Chris Seilboth, Elsa Dent. ................................................................. 11
Usability and Acceptability of a Palliative Care Mobile Intervention for Older Adults With Heart Failure and Caregivers: Observational Study

Jennifer Paola Villalobos¹, MS; Sheana Salyers Bull¹, MPH, PhD; Jennifer Dickman Portz², MSW, PhD

¹Colorado School of Public Health, University of Colorado, Aurora, CO, United States
²Division of General Internal Medicine, University of Colorado, Aurora, CO, United States

Corresponding Author:
Jennifer Paola Villalobos, MS
Colorado School of Public Health
University of Colorado
13055 E 17th Ave
Mail Stop #F802
Aurora, CO, 80045
United States
Phone: 1 303 551 3823
Email: jennifer.p.villalobos@gmail.com

Abstract

Background: Heart failure is a leading cause of death among older adults. Digital health can increase access to and awareness of palliative care for patients with advanced heart failure and their caregivers. However, few palliative care digital interventions target heart failure or patients’ caregivers, family, and friends, termed here as the social convoy. To address this need, the Social Convoy Palliative Care (Convoy-Pal) mobile intervention was developed to deliver self-management tools and palliative care resources to older adults with advanced heart failure and their social convoys.

Objective: The goal of the research was to test the acceptability and usability of Convoy-Pal among older adults with advanced heart failure and their social convoys.

Methods: Convoy-Pal includes tablet-based and smartwatch tools facilitating self-management and access to palliative care resources. Older adults and social convoy caregivers completed an acceptability and usability interview via Zoom, including open-ended questions and the Mobile Application Rating Scale: User Version (uMARS). Descriptive analysis was conducted to summarize the results of open-ended feedback and self-reported acceptability and usability.

Results: A total of 26 participants (16 older adults and 10 social convoy caregivers) participated in the interview. Overall, the feedback from users was good (uMARS mean 3.96/5 [SD 0.81]). Both older adults and social convoy caregivers scored information provided by Convoy-Pal the highest (mean 4.22 [SD 0.75] and mean 4.21 [SD 0.64], respectively). Aesthetics, functionality, and engagement were also perceived as acceptable (mean >3.5). Open-ended feedback resulted in 5 themes including improvements to goal setting, monitoring tools, daily check-in call feature, portal and mobile app, and convoy assessment.

Conclusions: Convoy-Pal was perceived as acceptable with good usability among older adults with heart failure and their social convoy caregivers. With good acceptability, Convoy-Pal may ultimately lead to increased access to palliative care resources and facilitate self-management among older adults with heart failure and their social convoy caregivers.

(JMIR Aging 2022;5(4):e35592) doi:10.2196/35592

KEYWORDS
mHealth; older adult; symptom; heart failure; palliative care; app; digital health; cardiology; heart; Convoy-Pal; mobile; tablet; smartwatch; adult; aging
**Introduction**

Heart failure (HF) is the 4th leading cause of death from heart disease in the United States and is most prevalent among individuals aged 65 years and older (ie, older adults). According to data from 2015 to 2018, 7.5% of males and 3.9% of females aged 60 to 79 years have HF [1]. The prevalence of HF continues to rise over time as the population ages [2]. By the year 2040, the number of older Americans is expected to nearly double to an estimated 80.8 million [3]. As the prevalence of HF increases, the need for palliative care amplifies. Palliative care can be beneficial for patients with HF as well as their caregivers, families, friends, and loved ones, referred to here as the social convoy [4]. Palliative care offers a support system to help the social convoy cope during the patient’s illness and effectively control distressing symptoms experienced by patients with HF [5]. In general, symptom control and good communication are basic palliative care principles highly recommended to improve the quality of life for patients with HF [5]. Although relatively underexplored, digital health [6] innovations (ie, telehealth, wearable devices, and mobile health [mHealth]) provide modern opportunities for patients and their social convoy to engage in palliative care [7-11].

Although there is a need, few studies focus on HF-specific mHealth in palliative care or mHealth supports for the social convoy. A systematic review of mHealth in palliative care reports that the primary uses of mobile apps are for biological and clinical monitoring (75% of the apps), disease self-management (64% of the apps), and therapeutic patient education (50% of the apps) [12]. One pilot in the review targets patients with HF and has found that the use of the HF mobile app improves self-care management [13]. Another study involving HF patients and their informal caregivers shows mHealth may decrease risk of HF exacerbations and improve caregiver communication [14]. While there are early indicators that patients and caregivers benefit from mHealth, providers also express enthusiasm about the potential of mHealth in palliative care [15-17]. Palliative care providers recommend digital health innovations in the areas of telehealth, client health records, and personal health tracking [17].

Quality testing in mHealth includes acceptability and usability as standard and essential in the field. Acceptability testing is usually completed first, followed by usability testing. This type of testing, for example, allows researchers to increase confidence that subsequent research on the efficacy of a tool produces outcomes that ensure null or negative outcomes are not due to poor tool function. Essentially, acceptability testing in mHealth assists with determining the level of meaningful engagement with the app; otherwise, if not engaging, the app will not be used, which may affect retention over time [18]. Usability testing, on the other hand, highlights the need to adapt the apps to users’ needs to create more usable tools [19] and ensure an app can be used the way it was intended by the specific audience for the tool [20].

Given the limited access to HF-specific palliative care mHealth, the Social Convoy Palliative Care Mobile Intervention, Convoy-Pal, was developed in response to a need for self-care strategies for both older adults with HF and their social convoys. Convoy-Pal was co-designed with older adults, caregivers, and health care providers [21] under the clinical guidelines establish by the National Coalition for Hospice and Palliative Care [22]. As a step in the co-designing process, this study was to test the acceptability of Convoy-Pal among older adults with HF and their caregivers.

**Methods**

**Convoy-Pal Platform**

The authors are researchers at the mHealth Impact Lab [23] who contract with Routinify, Inc, [24], a vendor that delivers the Convoy-Pal intervention. Routinify offers a variety of software and hardware tools that are publicly available; costs vary based on the tools provided and can range from US $50 to $100 per patient. In this case, Routinify assists with the delivery of the Convoy-Pal intervention to older adults and their social convoys. However, Routinify is only permitted to deliver Convoy-Pal in contract with the mHealth Lab and is not engaged in the clinical research (ie, they are not involved in the study instruments, data collection, management, analysis, or designing the protocol).

Convoy-Pal is designed with the following care domains: physical, psychological, social, spiritual, near end of life, ethical and legal, and knowledge about palliative care overall. Convoy-Pal includes a palliative care assessment with self-monitoring and resource tools for each domain. For example, the near end-of-life aspects of care (Figure 1) includes information regarding grief support and self-care and provides an opportunity for life review activities. This also includes resources on how to communicate unaddressed concerns and identify a support group for social support. Convoy-Pal tools and content are designed to be delivered via WellAssist by Routinify, Inc. WellAssist is a personal point-of-care app and associated internet-connected medical devices. The app’s core is based on behavioral modifications in line with the overall plan of care. The app is designed so that all members within the social convoy can access and use the Convoy-Pal intervention.
Ethics Approval and Considerations

Study procedures were approved by the Colorado Multiple Institutional Review Board (number 18-0973). All participants electronically consented to participate. Convoy-Pal collects assessments, including smartwatch-captured vital information, regarding mental health and overall well-being data for the older adult patient and caregivers. It is ethically sound to obtain consent from older adults to share their health information with their caregivers as we did for this study; however, older adult patients also have the option not to share their health information with their caregivers if desired.

Recruitment

Potential patient participants were identified from the UCHealth University of Colorado Hospital health system’s electronic medical record. Potential participants were aged at least 65 years at the time of recruitment and had been hospitalized at the UCHealth Hospital more than 2 times for HF in the year prior (January 2020-2021). Participants were currently living in their homes and receiving follow-up HF care. We mailed a study invitation letter with an opt-out contact option. Patients who did not opt out were then contacted by phone for recruitment and asked to self-identify social convoy caregivers.

Data Collection

Two research coordinators (JPV and IM) interviewed participants via Zoom to gain feedback on Convoy-Pal. Participants were exposed to sections of Convoy-Pal throughout the interview process, which lasted between 40 minutes to 1 hour. Participants were shown the Convoy-Pal hardware, which consisted of the tablet, watch, and charging station, and the web-based system portal and mobile app during the interview. Participants were also shown Convoy-Pal features such as goal setting and planning, monitoring options, daily check-in and calling features, convoy caregiver assessments, and palliative care resources. During the exposure to Convoy-Pal, participants completed a self-report acceptability measure and were asked to provide open-ended feedback.

For self-report acceptability, participants completed the Mobile Application Rating Scale: User Version (uMARS) survey [25]. The uMARS survey comprises 4 objective quality subscales—(1) engagement with the app, (2) functionality and users’ perceived functioning of the app, (3) aesthetics, and (4) users’ perception of the quality of the information [25]—to determine app quality mean score. uMARS has 2 optional subscales that can be used depending on the aims of the research. These 2 subscales are the app subjective quality scale, which can be reported as individual items, and the perceived impact scale, which obtains information on the knowledge, attitudes, and behavior change toward improving health behavior [25]. All items are assessed on a 5-point scale, with a uMARS score of 5 considered excellent while a score of 1 is inadequate [26]. The uMARS is shown to have high interrater reliability for evaluating the quality of mHealth apps on well-being, for example [25,27].

For open-ended feedback, the tablet and smartwatch were introduced and displayed over Zoom to the participants. We used a semistructured interview guide (Multimedia Appendix 1) to ask the participants questions and their opinions regarding the hardware, goal setting and planning, monitoring options, daily check-in and convoy calling options, portal and mobile app for convoy, convoy assessments, and convoy resources. Notes, recommendations, and opinions from participants were archived into Qualtrics (Qualtrics) [28], data management software, as the interview was being conducted. The final data captured were stored and saved in Qualtrics with their study ID numbers.

Data Analysis

The uMARS survey data was analyzed using Excel (Microsoft Corp) calculation mechanisms and descriptive frequencies including mean scores for both caregivers and patients. Once
all the interviews were complete, the qualitative data were moved from Qualtrics to NVivo 12 (QSR International) [29], a qualitative research software package, for analysis. A preliminary codebook was created, incorporating explicit domains from the interview guide (deductive themes) by a research assistant (JPV). A descriptive qualitative approach [30] was then used to identify themes and subthemes [31,32]. The codebook and final data interpretation were discussed in a group with all authors. No member checking was conducted.

**Results**

**Participants**

We recruited 26 participants (16 patients and 10 caregivers) from the University of Colorado Denver and its affiliate, the University of Colorado Hospital. Patients and convoy caregivers participated together or separately depending on the patient’s ability to participate in the interview. Patients were primarily males (9/16, 56%), White (14/16, 88%), and had a mean age of 76 (SD 5) years. Caregivers were predominantly female (7/10, 70%), White (7/10, 70%), and had a mean age of 71 (SD 10) years. Patients were married (12/16, 75%) and had a postgraduate degree (8/16, 50%), with 44% (7/16) having an income of US $30,000 or more and 82% (13/16) owning an iPhone, Android, or a regular or basic phone (Table 1). Similarly, most caregivers were married (8/10, 80%) and had a college or postgraduate degree (7/10, 70%), with 50% (5/10) making US $30,000 or more; 40% (4/10) of caregivers chose not to answer the question regarding their income. All of the caregivers owned an iPhone, Android, or a regular basic phone (Table 1). Due to small cell sizes, demographic categories were collapsed and are not reported in the table.

<table>
<thead>
<tr>
<th>Technology use</th>
<th>Patients, n (%) (n=16)</th>
<th>Caregivers, n (%) (n=10)</th>
<th>Total, n (%) (N=26)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell phone</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic phone: iPhone, Android, or regular or basic phone</td>
<td>13 (82)</td>
<td>10 (100)</td>
<td>23 (84)</td>
</tr>
<tr>
<td>I do not have a cell phone</td>
<td>1 (&lt;1)</td>
<td>0</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td>Did not respond</td>
<td>2 (13)</td>
<td>_a</td>
<td>2 (1)</td>
</tr>
<tr>
<td><strong>Digital activity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email</td>
<td>16 (100)</td>
<td>10 (100)</td>
<td>26 (100)</td>
</tr>
<tr>
<td>Look up information</td>
<td>16 (100)</td>
<td>10 (100)</td>
<td>26 (100)</td>
</tr>
<tr>
<td>Use social media</td>
<td>16 (100)</td>
<td>6 (60)</td>
<td>22 (84)</td>
</tr>
<tr>
<td>Post and share pictures or videos</td>
<td>15 (94)</td>
<td>9 (90)</td>
<td>24 (92)</td>
</tr>
<tr>
<td>Read or post comments</td>
<td>15 (94)</td>
<td>10 (100)</td>
<td>25 (96)</td>
</tr>
<tr>
<td>Play computer games</td>
<td>14 (88)</td>
<td>10 (100)</td>
<td>24 (92)</td>
</tr>
<tr>
<td>Video chat</td>
<td>16 (100)</td>
<td>10 (100)</td>
<td>26 (100)</td>
</tr>
<tr>
<td>Instant message or chat rooms</td>
<td>13 (82)</td>
<td>7 (70)</td>
<td>20 (76)</td>
</tr>
</tbody>
</table>

_aNot applicable._

**Acceptability**

**Mobile Application Rating Scale: User Version**

Overall, the acceptability feedback from users was good. The uMARS mean score was 4.00 (SD 0.78) among patients and 3.92 (SD 0.83) among caregivers, with an overall uMARS mean score of 3.96 (SD 0.81) among both groups (Table 2). Patients and caregivers showed the most concordance with Section D: information scale and the most discordance with Section C: aesthetics (Table 2). Further description of the mean, standard deviation, and minimum and maximum values for the subscales of the uMARS are provided in Multimedia Appendix 2.

Examining uMARS domain scores individually, we found that patients gave Section D: information the highest rating (mean 4.22, SD 0.75), followed by Section C: aesthetics (mean 4.13, SD 0.73), Section B: functionality (mean 3.87, SD 0.85), and Section A: engagement (mean 3.80, SD 0.79). Patients scored the app’s subjective quality scale a mean of 4.01 (SD 0.70) and the perceived impact of the app on the user’s knowledge, attitudes, and intentions related to the target health behavior a 3.64 (SD 0.96). Similarly, caregivers scored Section D: information the highest (mean 4.21, SD 0.64), followed by Section C: aesthetics (mean 3.89, SD 0.72), Section B: functionality (mean 3.82, SD 1.0), and Section A: engagement (mean 3.77, SD 0.96). The app subjective quality scale was rated mean 3.56 (SD 1.23) and perceived impact was rated mean 3.13 (SD 1.20) among caregivers.
Table 2. Mean, standard deviation, and range values for the subscales of the uMARS (Mobile Application Rating Scale: User Version).

<table>
<thead>
<tr>
<th>Section</th>
<th>Patients Mean (SD)</th>
<th>Range</th>
<th>Caregivers Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: engagement</td>
<td>3.80 (0.79)</td>
<td>2.25-5.00</td>
<td>3.77 (0.96)</td>
<td>2.40-5.00</td>
</tr>
<tr>
<td>B: functionality</td>
<td>3.87 (0.85)</td>
<td>2.00-5.00</td>
<td>3.82 (1.00)</td>
<td>2.50-5.00</td>
</tr>
<tr>
<td>C: aesthetics</td>
<td>4.13 (0.73)</td>
<td>3.33-5.00</td>
<td>3.89 (0.72)</td>
<td>3.00-5.00</td>
</tr>
<tr>
<td>D: information</td>
<td>4.22 (0.75)</td>
<td>2.50-5.00</td>
<td>4.21 (0.64)</td>
<td>3.50-5.00</td>
</tr>
<tr>
<td>Total</td>
<td>4.00 (0.78)</td>
<td><em>a</em></td>
<td>3.92 (0.83)</td>
<td>_</td>
</tr>
</tbody>
</table>

*aNot applicable.*

**Open-Ended Feedback**

Five main themes were identified after receiving open-ended feedback: goal setting, monitoring tools, daily check-in call feature, portal and mobile app, and convoy assessment. Representative quotes for themes and additional subthemes are reported in Table 3.

**Goal Setting**

Participants expressed the need for the goal-setting section to provide realistic and obtainable goals. For example, it was expressed that goal setting should be addressed monthly, not weekly. Additionally, participants expressed that they would like an option to add a comment box to include other action items and or commentary for their goals.

**Monitoring Tools**

Participants agreed that the monitoring tools were helpful for people with HF and other chronic conditions. Feedback to improve the monitoring tools included the addition of other features, such as feedback prompting concerning vitals, electrocardiogram measures, fall detection, stroke indicators, and heart palpitation monitoring.

**Reminder and Call Feature**

Most of the participants liked the daily check-in and call feature. One participant said “...the feature is good for people who live alone and want to keep in contact via FaceTime with their loved ones” (72-year-old participant). A few participants who disliked the daily check-in feature expressed that some patients might feel burdened by the frequency of check-ins. Others indicated the feature was redundant as they could schedule reminders and calls through their personal phone instead. One suggestion included designing a checklist of all of the medication a person takes and getting notifications on those specific mediations versus just getting a general message.

**Portal and Mobile App**

The majority of the participants liked the portal and mobile app. The participants appreciated that the portal, charts, and layout of the mobile app were clear and concise. Participants also liked the opportunity to share access to personal data with family members. Feedback from 2 participants included adding thresholds to the graphs to determine who, such as a provider or family member, should be consulted, and adding instructions on who to call with concerns.

**Social Convoy Assessment**

Participants provided many recommendations when asked their opinions about convoy assessment. Many participants were hesitant about the caregiver assessments due to time, burden, and specific assessment topics. For example, in the domain of emotional assessment, a participant said “if [a caregiver] is [in] crisis mode they will not fill out the questions...this is not beneficial for patients who need extra help and support” (Table 3). Another participant expressed that “they would not answer those questionnaires truthfully because they were not raised to share emotions growing up” (Table 3). Additionally, a
participant expressed that some assessments should be addressed in person and not via the tablet. Other themes that arose from the feedback included ensuring assessments are HIPAA (Health Insurance Portability and Accountability Act) compliant and appropriate language is used. For example, one participant responded, “The ‘I feel sad’ language might not be appropriate for people because they are not readily going to admit that they are sad” (71-year-old participant). Participants also suggested that the assessments should not take too long to complete.

Discussion

Principal Findings
Convoy-Pal is designed to increase access to palliative care resources and self-management in the setting of HF. Acceptability testing is essential because it results in a better quality product. In this acceptability and usability study, the Convoy-Pal is considered acceptable and a good quality app, based on the uMARS scores among older adults with HF and their caregivers. Older adult patients and caregivers also provided recommendations for improving Convoy-Pal, which included adding comment sections, designing a checklist for medications, including thresholds on graphs for interpretation, and adding features such as fall detection. Based on this feedback, authors will update and continue to assess Convoy-Pal for usability and feasibility.

Although there are few high-quality HF mobile apps [33] assessed for acceptability, functionality, and efficacy [33,34], our findings are supported by other empirical studies. For example, palliative care patients found a mobile mortality risk tool acceptable to use [35]. Additionally, using wearables for monitoring palliative care was also feasible [36,37], a tool that Convoy-Pal offers with its smartwatch. Similar to the feedback provided for Convoy-Pal, a commentary article, systematic meta-review, and qualitative study [15,38,39] report the need to track relevant information, receive education pertinent to health for older adults, and provide information sharing such as medication use.

Aside from HF apps specifically, digital health interventions overall have the potential to improve the accessibility and effectiveness in palliative care, as reported by a recent systematic meta-review [38]. Palliative care is one area where technologies are increasingly being deployed. Although leveraging existing resources for palliative care is one approach, mHealth interventions targeting palliative care enable patients increased access to this resource without spending time or traveling to locations [40]. mHealth palliative care allows older adults to participate in and govern their care. For example, they do this by self-reporting symptoms and needs, which improves communication with providers and caregivers [38,41,42]. Traditional palliative care resources do not provide a self-governing element in this unique way.

Additionally, HF mobile interventions rarely target the social convoy or palliative care domains. Therefore, Convoy-Pal would contribute to the advancement of palliative care and HF mHealth while also advancing a team approach to information sharing and targeting family- and caregiver-specific issues [43]. Convoy-Pal has the potential to support older adults with HF and their social convoy in the management of physical, psychosocial, and spiritual concerns.

Limitations
First, due to university and state COVID-19 restrictions, research assistants were not able to meet with older adult and caregiver participants to physically interact with Convoy-Pal on the Routinify tablets or complete assessments in person. Researchers therefore collected acceptability and usability data by displaying Convoy-Pal and all its features remotely to participants for about 1 hour through Zoom. Second, the uMARS survey, for this reason, was modified by our team to reflect the following 2 optional responses for all subscales of the survey: (1) “Optional: Missing due to lack of time with app” and (2) “Optional: Did not feel comfortable answering.” If the participant did not feel comfortable answering the uMARS questions due to their belief that there was not enough time to explore Convoy-Pal, then they could select either optional response. The minor modifications made to the uMARS had not previously been tested and therefore may have reduced the validity of the original items. Physical interaction with the hardware may have yielded additional user feedback. Last, the study was further limited by small sample size and a single health system as well as lack of diversity representative of the local community. Acceptability and usability of Convoy-Pal may differ in other regional areas and varying access to health care.

Conclusion
HF is a leading cause of death in the United States, and mHealth provides opportunities for patients and their social convoy to participate in palliative care. In our study, 16 older patients and 10 caregivers were interviewed and asked to complete a uMARS assessment and provide open-ended feedback. Overall, older patients and their caregivers perceived Convoy-Pal as acceptable with good usability. Although in-person usability testing is needed, Convoy-Pal was perceived acceptable and may ultimately increase access to palliative care resources and facilitate self-management among older adults with HF and their caregivers.

Acknowledgments
The authors thank Routinify, Inc, for its assistance with the development and assessment of the Convoy-Pal. We also thank research assistant Irazema Mino for assistance conducting some qualitative interviews. This study was funded by grant K76AG059934 from the National Institute on Aging.
Data Availability
The data sets generated and analyzed during this study are not yet publicly available. The full study protocol and data sets will be shared to a public data repository upon the conclusion of the Convoy-Pal feasibility study.

Authors’ Contributions
JDP, as principal investigator, supervised the project and manuscript, while JPV collected the data, conducted the analysis, and drafted the manuscript. SSB provided general feedback and suggestions on the direction and context of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Convoy-Pal Acceptability Interview Guide.

[DOCX File, 26 KB - aging_v514e35592_app1.docx ]

Multimedia Appendix 2
Mean, standard deviation, and minimum and maximum values for the scales and subscales of the uMARS (Mobile Application Rating Scale: User Version).

[DOCX File, 17 KB - aging_v514e35592_app2.docx ]

References


29. NVivo. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home


**Abbreviations**

- **Convoy-Pal**: Social Convoy Palliative Care
- **HF**: heart failure
- **mHealth**: mobile health
- **uMARS**: Mobile Application Rating Scale: User Version

©Jennifer Paola Villalobos, Sheana Salyers Bull, Jennifer Dickman Portz. Originally published in JMIR Aging (https://aging.jmir.org), 06.10.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Aging, is properly cited. The complete bibliographic information, a link to the original publication on https://aging.jmir.org, as well as this copyright and license information must be included.
Evolving Hybrid Partial Genetic Algorithm Classification Model for Cost-effective Frailty Screening: Investigative Study

John Oates1*, PhD; Niusha Shafiabady2*, BEng (Hon), MSc, PhD; Rachel Ambagtsheer3, PhD; Justin Beilby3, MD, PhD; Chris Seiboth4, BA; Elsa Dent3, PhD

1Torrens University, Ultimo, Australia
2College of Engineering, Information Technology and Environment, Charles Darwin University, Haymarket, NSW, Australia
3Torrens University, Adelaide, Australia
4Lutheran Services, Milton, Australia
*these authors contributed equally

Corresponding Author:
Niusha Shafiabady, BEng (Hon), MSc, PhD
College of Engineering, Information Technology and Environment
Charles Darwin University
815 George Street
Haymarket, NSW, 2000
Australia
Phone: 61 80474147
Email: niusha.shafiabady@cdu.edu.au

Abstract

Background: A commonly used method for measuring frailty is the accumulation of deficits expressed as a frailty index (FI). FIs can be readily adapted to many databases, as the parameters to use are not prescribed but rather reflect a subset of extracted features (variables). Unfortunately, the structure of many databases does not permit the direct extraction of a suitable subset, requiring additional effort to determine and verify the value of features for each record and thus significantly increasing cost.

Objective: Our objective is to describe how an artificial intelligence (AI) optimization technique called partial genetic algorithms can be used to refine the subset of features used to calculate an FI and favor features that have the least cost of acquisition.

Methods: This is a secondary analysis of a residential care database compiled from 10 facilities in Queensland, Australia. The database is comprised of routinely collected administrative data and unstructured patient notes for 592 residents aged 75 years and over. The primary study derived an electronic frailty index (eFI) calculated from 36 suitable features. We then structurally modified a genetic algorithm to find an optimal predictor of the calculated eFI (0.21 threshold) from 2 sets of features. Partial genetic algorithms were used to optimize 4 underlying classification models: logistic regression, decision trees, random forest, and support vector machines.

Results: Among the underlying models, logistic regression was found to produce the best models in almost all scenarios and feature set sizes. The best models were built using all the low-cost features and as few as 10 high-cost features, and they performed well enough (sensitivity 89%, specificity 87%) to be considered candidates for a low-cost frailty screening test.

Conclusions: In this study, a systematic approach for selecting an optimal set of features with a low cost of acquisition and performance comparable to the eFI for detecting frailty was demonstrated on an aged care database. Partial genetic algorithms have proven useful in offering a trade-off between cost and accuracy to systematically identify frailty.

(JMIR Aging 2022;5(4):e38464) doi:10.2196/38464

KEYWORDS
machine learning; frailty screening; partial genetic algorithms; SVM; KNN; decision trees; frailty; algorithm; cost; model; index; database; ai; ageing; adults; older people; screening; tool
**Introduction**

Genetic algorithms (GA) are a general-purpose computational optimization method inspired by the evolution mechanism in nature. They are one of the most popular metaheuristic search algorithms and have been used for variety of applications, including synthetic data generation, feature selection, and to solve complex equations [1]. In this study, genetics algorithms have been applied to identify features that offer a suitable trade-off between cost and accuracy.

Within the context of global population aging, the number of older people who will live a significant proportion of their lives with frailty is growing rapidly [2]. Frailty is problematic for older people and the societies in which they live due to the elevated risks associated with the syndrome and terms poor health outcomes [3] and additional use of health and aged care services [4-7], leading to inflated health care costs [8-10]. However, emerging research suggests that frailty is a highly dynamic [11,12] and potentially modifiable state with appropriate intervention [13,14]. Screening for early detection is proposed to increase the likelihood that the worst impacts of frailty can be lessened [4,15,16].

There are 2 main approaches to identifying frailty: the frailty phenotype (FP) and the frailty index (FI) [17]. However, these established approaches have known drawbacks, requiring significant time investment, face-to-face interaction, and specific data items to be collected [18]. Recently, an electronic frailty index (eFI) was proposed [19] that has the potential to achieve greater efficiencies over face-to-face models when applied to administrative data sets, but the need to ensure a minimum set of items adhering to prespecified criteria remains a barrier to implementation. For example, previous research has shown that although it is possible to calculate and construct an eFI based on an aged care administrative data set, a significant proportion of the items require manual calculation to ensure accuracy and improve quality [20]. Clearly, it would be preferable to identify automated techniques capable of delivering comparable accuracy and quality but with greater efficiency. Consequently, this study aimed to apply a sophisticated genetic algorithm technique to identify an optimal predictor of the calculated eFI.

**Methods**

**Study Design, Participants, and Setting**

This retrospective study utilized a data set previously compiled [21] from the administrative database of 10 residential aged care facilities located in Queensland, Australia. Participants were included in the study if they were aged 75 years or older and had completed an Aged Care Funding Instrument (ACFI) assessment within the previous 3 years.

**Ethical Considerations**

A waiver of consent for the initial study was obtained from the Human Research Ethics Committee of Torrens University Australia (application H11/19), which declared the study exempt under National Statement 5.1.22 (secondary use of deidentified administrative data) due to the pragmatic nature of the study. Because this is a secondary study of the same data, the approval extends to this study. Moreover, this study adheres to the Australian National Statement on Ethical Conduct in Human Research.

**Frailty Outcome Measure**

An eFI was previously calculated for this data [21] based on a formulation originally specified by Clegg et al [22]. Care was taken to ensure the included deficits adhered to the criteria recommended by Searle and colleagues [23], which resulted in 32 of the 35 deficits being extracted from unstructured patient notes and only 3 being derived from the ACFI data. The binary frailty classification was derived using a threshold of 0.21 (ie, frailty defined as >0.21) [24].

**Screening Test Construction**

Genetic algorithms are an optimization technique [1] applied in machine learning to filter a set of features that are used to construct a classification model. During training, a classification algorithm is tuned on a training set, and the success of attaining a generalized predictive algorithm is then verified by measuring the classification errors in the test set.

Genetic algorithms leverage the observation that classification models often perform better when they are trained on a subset of the available features. Which subset of features to use, however, is not obvious. Genetic algorithms start with a population of randomly generated subsets of features, or chromosomes, that are all independently used to generate classification models. The chromosomes from the population that generated the best performing models are allowed to combine, or breed, to form a new generation of the population, while the worst performing ones are removed completely. The process continues until either a predefined number of generations have been trained or the performance of the models has plateaued. Once training is complete, the best-performing model is deployed using only the naturally selected subset of the available features.

While genetic algorithms are good at selecting an optimal subset of features, they select the features based on maximizing the classification accuracy of a generated model. The cost of acquiring the various features is not factored into the choice of features, even if the performance of less expensive features is close to that of their more expensive counterparts. In this study, the cost of a feature is the combination of the effort, monetary cost, and patient risk involved in capturing the values. We want to minimize the number of expensive features chosen to form the model but allow as many low-cost features to be used as is necessary to gain acceptable performance of the model.

To achieve the inclusion of low-cost features in the classification model, the standard genetic algorithm training configuration illustrated in Figure 1 is modified as illustrated in Figure 2.
This modification is performed every time a model is trained for every member of the population trialed by the genetic algorithm. When the genetic algorithm trains a model, it passes a subset of the available training records to the classification model’s training algorithm. The low-cost feature values for each record need to be added to the selected training records before commencing the training. The genetic algorithm trains the classification model for each chromosome multiple times with different subsets of the training records and determining the performance of each model using records not used in training that instance. As with the training records, the low-cost features need to be added to the records used to determine a model’s performance. The performance of the chromosome is calculated as the average performance of all the models built from different subsets of the training records. This process is called n-fold cross validation, where n is the number of models built. In this study, 3-fold cross validation was used because it ensured a good balance between performance and the time it took to build the models.

Four types of classification models were optimized using partial genetic algorithms: logistic regression, support vector machines, random forest, and decision trees. These algorithms are popular choices for classification because they have proven successful in generating generalized models for a wide range of applications [20]. Logistic regression is a statistical modeling technique whereby a linear combination of the input features is found during training, which models the logarithm of the odds that a binary outcome is in the true state. A support vector machine (SVM) aims to learn a multidimensional hyperplane that separates the set of records given to it for training. Predictions are made by placing the candidate record in the same multidimensional classification space and determining which side of the hyperplane it maps to. SVM was developed in the 1990s and has since enjoyed success in many real-world applications.
applications, including pattern recognition [25], text classification [26], and bioinformatics. Decision trees employ a divide and conquer strategy. A tree is formed of nodes, and each node performs a comparison of a single input feature and a threshold if the variable is continuous or a state if the feature is discrete. The outcome of the comparison determines the choice of the next node, which either performs a new comparison or terminates the tree with a given classification. During training, the set of training records are used to find comparisons at each node that gain the most information by reducing entropy in the outcomes by the greatest amount. Subsequent training predictions are made by feeding records into the root node and determining the classification of the terminating node where the record exits the tree. Random forest is a meta form of decision trees, where the output is determined by a vote between many trees. The trees are built using different methods to ensure they are not replicas of each other.

The software was written in Python and the models were built using the sklearn module (version 1.0.2) and the genetic_selection module from sklearn-genetic (version 0.5.1).

**Results**

**Model Generation**

Of the 69 features considered, 34 were extracted directly from the ACFI assessment and 35 were the values used to calculate the eFI. Two of the ACFI features, Psychogeriatric Assessment Scales (PAS) score and Cornell Scale, were excluded as they had a high percentage of missing values (PAS score 36%, Cornell Scale 42%). The remaining 32 ACFI assessment features had no missing values and were categorized as low cost of acquisition features. Of the 35 features used to calculate the eFI, 32 were extracted by an automated search for key words in the unstructured patient notes, followed by manual inspection and verification by a clinician. These were categorized as having a high cost of acquisition. The remaining 3 features used to calculate the eFI were direct combinations of ACFI features. As the calculation of these features could be fully automated, they were included with the low-cost features. A total of 4 sets of low-cost features were considered: (1) ACFI features + the low-cost eFI features; (2) the low-cost eFI features; (3) no low-cost features; and (4) a set of features chosen from the low-cost features using genetic algorithms. A different set was found for each of the classification algorithms.

Sixteen scenarios were trialed, comprising each of the aforementioned 4 sets of low-cost features for each of the 4 classification algorithms. For each scenario, the partial genetic algorithm was used to optimize the classification algorithm with different limits placed on the number of high-cost features. The limits were varied sequentially from 1 to 32, which was the number of candidate high-cost features. The performance of each of the 32 algorithms generated for each scenario were plotted on a single graph. The graphs for each scenario are plotted in Figures 3-6.

When comparing the graphs for each classification model, logistic regression outperformed decision trees in every scenario and SVM and random forest in almost all scenarios. Tables 1-3 demonstrate the numeric comparison of the 16 scenarios when 5, 10, and 15 of the high cost of acquisition features were used.

The option of “No low-cost” features was provided to determine how much predictive value the low-cost features were adding to the classification. As expected, this option performed the worst for all the classification algorithms, confirming that the low-cost features were adding value. Next, models were built using only the 3 low-cost eFI features as fixed features. This improved the accuracy of the logistic regression algorithm to 97% when almost all the eFI features were included (Table 4). Although this is a good outcome, a model built using so many of the high-cost features was not the goal of this study.

A genetic algorithm works by selecting an optimal subset of all the features made available to it. This characteristic was the motivation behind building a version of the models in 2 stages. In the first stage, a standard, nonpartial, genetic algorithm was used on the low-cost features to find an optimal combination. These models performed so poorly (Table 5) that they could not be used without further improvement. The combination of features used to generate these models (Multimedia Appendices 1-3) was then employed as the fixed features in the partial genetic algorithm during the second stage. The models in the second stage performed surprisingly poorly, showing no difference from the models built without any low-cost features, regardless of the classification model used.

Using all the low-cost features in a partial genetic algorithm yielded the best overall results and matched the 97% accuracy achieved by the models that used the low-cost eFI features when the model was able to select most of the high-cost eFI features. At 10 features, however, the extra low-cost features allowed the algorithm to increase its sensitivity from 82.7% to 89.3% and specificity from 81.7% to 86.7%.
Figure 3. Logistic regression optimized with a partial genetic algorithm. ACFI: Aged Care Funding Instrument; EFI: electronic frailty index; GA: Genetic algorithm; LR: logistic regression; npa: negative percent agreement; ppa: positive percent agreement.
Figure 4. Support vector machine optimized with a partial genetic algorithm. ACFI: Aged Care Funding Instrument; EFI: electronic frailty index; GA: Genetic algorithm; npa: negative percent agreement; ppa: positive percent agreement; SVM: support vector machine.
Figure 5. Decision tree optimized with a partial genetic algorithm. ACFI: Aged Care Funding Instrument; DT: decision tree; EFI: electronic frailty index; GA: Genetic algorithm; npa: negative percent agreement; ppa: positive percent agreement.
**Figure 6.** Random forest optimized with a partial genetic algorithm. ACFI: Aged Care Funding Instrument; EFI: electronic frailty index; GA: Genetic algorithm; npa: negative percent agreement; ppa: positive percent agreement; RF: random forest.
Table 1. Performance of the 12 scenarios with 5 high-cost features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPA(^a)</th>
<th>NPA(^b)</th>
<th>Accuracy</th>
<th>F1(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACFI(^d) + low-cost eFI(^e)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>76</td>
<td>75</td>
<td>71.4</td>
<td>79.2</td>
<td>75.6</td>
<td>73.2</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>76</td>
<td>61.7</td>
<td>67.3</td>
<td>71.3</td>
<td>69.6</td>
<td>64.3</td>
</tr>
<tr>
<td>Decision tree</td>
<td>73.3</td>
<td>63.3</td>
<td>65.5</td>
<td>71.4</td>
<td>68.8</td>
<td>64.4</td>
</tr>
<tr>
<td>Random forest</td>
<td>80</td>
<td>71.7</td>
<td>74.1</td>
<td>77.9</td>
<td>76.3</td>
<td>72.9</td>
</tr>
<tr>
<td><strong>Low-cost eFI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>80</td>
<td>81.7</td>
<td>76.6</td>
<td>84.5</td>
<td>80.7</td>
<td>79</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>74.7</td>
<td>66.7</td>
<td>67.8</td>
<td>73.7</td>
<td>71.1</td>
<td>67.2</td>
</tr>
<tr>
<td>Decision tree</td>
<td>76</td>
<td>53.3</td>
<td>64.0</td>
<td>67.1</td>
<td>65.9</td>
<td>58.2</td>
</tr>
<tr>
<td>Random forest</td>
<td>72</td>
<td>70</td>
<td>66.7</td>
<td>75</td>
<td>71.1</td>
<td>68.3</td>
</tr>
<tr>
<td><strong>No low-cost features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>78.7</td>
<td>66.7</td>
<td>71.4</td>
<td>74.7</td>
<td>73.3</td>
<td>69</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>82.6</td>
<td>55</td>
<td>71.2</td>
<td>69.7</td>
<td>70.4</td>
<td>62.2</td>
</tr>
<tr>
<td>Decision tree</td>
<td>84</td>
<td>55</td>
<td>73.3</td>
<td>70</td>
<td>71.1</td>
<td>62.9</td>
</tr>
<tr>
<td>Random forest</td>
<td>76</td>
<td>65</td>
<td>68.4</td>
<td>73.1</td>
<td>71.1</td>
<td>66.7</td>
</tr>
<tr>
<td><strong>Genetically selected low-cost features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>80</td>
<td>66.7</td>
<td>72.7</td>
<td>75</td>
<td>74.1</td>
<td>69.6</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>82.7</td>
<td>58.3</td>
<td>72.9</td>
<td>71.3</td>
<td>71.9</td>
<td>64.8</td>
</tr>
<tr>
<td>Decision tree</td>
<td>76</td>
<td>68.3</td>
<td>69.5</td>
<td>75</td>
<td>77.2</td>
<td>68.9</td>
</tr>
<tr>
<td>Random forest</td>
<td>81.3</td>
<td>75</td>
<td>76.3</td>
<td>80.3</td>
<td>78.5</td>
<td>75.6</td>
</tr>
</tbody>
</table>

\(^a\)PPA: positive percent agreement.
\(^b\)NPA: negative percent agreement.
\(^c\)F1: F-score.
\(^d\)ACFI: Aged Care Funding Instrument.
\(^e\)eFI: electronic frailty index.
### Table 2. Performance of the 12 scenarios with 10 high-cost features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPA&lt;sup&gt;a&lt;/sup&gt;</th>
<th>NPA&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Accuracy</th>
<th>F1&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACFI&lt;sup&gt;d&lt;/sup&gt; + low-cost eFI&lt;sup&gt;e&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>89.3</td>
<td>86.7</td>
<td>86.7</td>
<td>89.3</td>
<td>88.1</td>
<td>86.7</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>85.3</td>
<td>80.0</td>
<td>81.4</td>
<td>84.2</td>
<td>83</td>
<td>80.7</td>
</tr>
<tr>
<td>Decision tree</td>
<td>65.3</td>
<td>61.7</td>
<td>58.7</td>
<td>68.1</td>
<td>63.7</td>
<td>60.2</td>
</tr>
<tr>
<td>Random forest</td>
<td>80</td>
<td>70</td>
<td>73.9</td>
<td>76.9</td>
<td>75.6</td>
<td>71.8</td>
</tr>
<tr>
<td><strong>Low-cost eFI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>82.7</td>
<td>81.7</td>
<td>79</td>
<td>84.9</td>
<td>82.2</td>
<td>80.3</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>81.3</td>
<td>81.7</td>
<td>77.8</td>
<td>84.7</td>
<td>81.5</td>
<td>79.7</td>
</tr>
<tr>
<td>Decision tree</td>
<td>81.3</td>
<td>50</td>
<td>68.2</td>
<td>67</td>
<td>67.4</td>
<td>57.7</td>
</tr>
<tr>
<td>Random forest</td>
<td>84</td>
<td>63.3</td>
<td>76</td>
<td>74.1</td>
<td>74.8</td>
<td>69.1</td>
</tr>
<tr>
<td><strong>No low-cost features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>72</td>
<td>86.7</td>
<td>71.2</td>
<td>87.1</td>
<td>78.5</td>
<td>78.2</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>77.3</td>
<td>78.3</td>
<td>73.4</td>
<td>81.7</td>
<td>77.8</td>
<td>75.8</td>
</tr>
<tr>
<td>Decision tree</td>
<td>81.3</td>
<td>70</td>
<td>75</td>
<td>77.2</td>
<td>76.3</td>
<td>72.4</td>
</tr>
<tr>
<td>Random forest</td>
<td>80</td>
<td>66.7</td>
<td>72.7</td>
<td>75</td>
<td>74.1</td>
<td>69.6</td>
</tr>
<tr>
<td><strong>Genetically selected low-cost features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>82.6</td>
<td>80</td>
<td>78.6</td>
<td>83.8</td>
<td>81.5</td>
<td>79.3</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>78.7</td>
<td>78.3</td>
<td>74.6</td>
<td>81.9</td>
<td>78.5</td>
<td>76.4</td>
</tr>
<tr>
<td>Decision tree</td>
<td>77.3</td>
<td>60</td>
<td>67.9</td>
<td>70.7</td>
<td>69.6</td>
<td>63.7</td>
</tr>
<tr>
<td>Random forest</td>
<td>78.7</td>
<td>66.7</td>
<td>71.4</td>
<td>74.7</td>
<td>73.3</td>
<td>69</td>
</tr>
</tbody>
</table>

<sup>a</sup>PPA: positive percent agreement.

<sup>b</sup>NPA: negative percent agreement.

<sup>c</sup>F1: F-score.

<sup>d</sup>ACFI: Aged Care Funding Instrument.

<sup>e</sup>eFI: electronic frailty index.
### Table 3. Performance of the 12 scenarios with 15 high-cost features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPA&lt;sup&gt;a&lt;/sup&gt;</th>
<th>NPA&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Accuracy</th>
<th>F1&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACFI&lt;sup&gt;d&lt;/sup&gt; + low-cost eFI&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>85.3</td>
<td>85.0</td>
<td>82.3</td>
<td>87.7</td>
<td>85.2</td>
<td>83.6</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>84.0</td>
<td>86.7</td>
<td>81.3</td>
<td>88.7</td>
<td>85.1</td>
<td>83.9</td>
</tr>
<tr>
<td>Decision tree</td>
<td>69.3</td>
<td>61.7</td>
<td>61.7</td>
<td>69.3</td>
<td>65.9</td>
<td>61.7</td>
</tr>
<tr>
<td>Random forest</td>
<td>84.0</td>
<td>66.7</td>
<td>76.9</td>
<td>75.9</td>
<td>76.3</td>
<td>71.4</td>
</tr>
<tr>
<td>Low-cost eFI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>85.3</td>
<td>81.7</td>
<td>81.7</td>
<td>85.3</td>
<td>83.7</td>
<td>81.7</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>86.7</td>
<td>81.7</td>
<td>83.1</td>
<td>85.5</td>
<td>84.4</td>
<td>82.4</td>
</tr>
<tr>
<td>Decision tree</td>
<td>76.0</td>
<td>58.3</td>
<td>66.0</td>
<td>69.5</td>
<td>68.1</td>
<td>61.9</td>
</tr>
<tr>
<td>Random forest</td>
<td>86.7</td>
<td>65.0</td>
<td>79.6</td>
<td>75.6</td>
<td>77.0</td>
<td>71.6</td>
</tr>
<tr>
<td>No low-cost features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>80.0</td>
<td>83.3</td>
<td>76.9</td>
<td>85.7</td>
<td>81.5</td>
<td>80.0</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>73.3</td>
<td>75.0</td>
<td>69.2</td>
<td>78.6</td>
<td>74.1</td>
<td>72.0</td>
</tr>
<tr>
<td>Decision tree</td>
<td>78.7</td>
<td>55.0</td>
<td>67.3</td>
<td>68.6</td>
<td>68.1</td>
<td>60.6</td>
</tr>
<tr>
<td>Random forest</td>
<td>77.3</td>
<td>76.6</td>
<td>73.0</td>
<td>80.6</td>
<td>77.0</td>
<td>74.8</td>
</tr>
<tr>
<td>Genetically selected low-cost features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>81.3</td>
<td>80.0</td>
<td>77.4</td>
<td>83.5</td>
<td>80.7</td>
<td>78.7</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>80.0</td>
<td>76.7</td>
<td>75.4</td>
<td>81.1</td>
<td>78.5</td>
<td>76.0</td>
</tr>
<tr>
<td>Decision tree</td>
<td>69.3</td>
<td>61.7</td>
<td>61.7</td>
<td>69.3</td>
<td>65.9</td>
<td>61.7</td>
</tr>
<tr>
<td>Random forest</td>
<td>84.0</td>
<td>66.7</td>
<td>76.9</td>
<td>75.9</td>
<td>76.3</td>
<td>71.4</td>
</tr>
</tbody>
</table>

<sup>a</sup>PFA: positive percent agreement.<br><sup>b</sup>NPA: negative percent agreement.<br><sup>c</sup>F1: F-score.<br><sup>d</sup>ACFI: Aged Care Funding Instrument.<br><sup>e</sup>eFI: electronic frailty index.

### Table 4. Performance of models based on all features.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPA&lt;sup&gt;a&lt;/sup&gt;</th>
<th>NPA&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Accuracy</th>
<th>F1&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR&lt;sup&gt;d&lt;/sup&gt;</td>
<td>97.3</td>
<td>96.7</td>
<td>96.7</td>
<td>97.3</td>
<td>97.0</td>
<td>96.7</td>
</tr>
<tr>
<td>SVM&lt;sup&gt;e&lt;/sup&gt;</td>
<td>86.7</td>
<td>95.0</td>
<td>85.1</td>
<td>95.6</td>
<td>90.4</td>
<td>89.8</td>
</tr>
<tr>
<td>DT&lt;sup&gt;f&lt;/sup&gt;</td>
<td>76.0</td>
<td>63.3</td>
<td>67.9</td>
<td>72.1</td>
<td>70.4</td>
<td>65.5</td>
</tr>
<tr>
<td>RF&lt;sup&gt;g&lt;/sup&gt;</td>
<td>88.0</td>
<td>75.0</td>
<td>83.3</td>
<td>81.5</td>
<td>82.2</td>
<td>78.9</td>
</tr>
</tbody>
</table>

<sup>a</sup>PFA: positive percent agreement.<br><sup>b</sup>NPA: negative percent agreement.<br><sup>c</sup>F1: F-score.<br><sup>d</sup>LR: logistic regression.<br><sup>e</sup>SVM: support vector machine.<br><sup>f</sup>DT: decision tree.<br><sup>g</sup>RF: random forest.
between the number of features and the performance of the derived models to be determined.

This study found that if a genetic algorithm was permitted to choose any number of features from all the available features, regardless of their cost, it most frequently chose subsets that only included high-cost features. This motivated the development of the previously mentioned partial genetic algorithm, which forced the algorithm to include low-cost features as well. However, this raises the question of whether the low-cost features add any value at all. To answer this question, the results include both a fixed set that had no low-cost features and a set that included only the low-cost features used to calculate the EFI. Considering logistic regression models with 10 high-cost features, including all the low-cost features, yielded an improvement of 17% in sensitivity (89% versus 72%). This combination did not compromise specificity, which remained stable (87%) and is comparable to the scenario with no low-cost features. This improvement is significant and possibly represents the difference between a clinically useful screening test and one that is inadequate. Even if the comparison is made between models built on all the low-cost features and those that include only low-cost features used in the EFI calculation, there is a 6% improvement in sensitivity (89% versus 83%) and 5% in specificity (87% versus 82%).

Although the partial genetic algorithm–built models with 10 high-cost features use less than a third of all the high-cost features, they still require those 10 features to be extracted by screening patient notes. Recent advances in natural language processing (NLP) show promise for automating this extraction process. It is plausible that NLP could extract all the features required to calculate the EFI, but this would require a much larger data set than the one used in this study. In the meantime, the cost of acquisition of at least 10 features from every patient record remains the cost of implementing a screening test on any database similar to ours that contains an ACFI assessment and unstructured patient notes.

Partial genetic algorithms can be used to derive classification models from any database where the cost of acquisition of some parameters is higher than it is for others. Although they have been demonstrated in this study on an aged care database to

### Discussion

#### Principal Findings

With AI techniques, cost-effective screening tests for frailty are possible for aged care databases that contain an ACFI assessment and unstructured patient notes. This study has shown that the ACFI assessment alone does not provide sufficient information to determine if a patient is frail. However, when ACFI data are augmented by as few as 10 additional features, an AI model can be derived that performs well enough to be used as a screening test. This means in clinical practice is that older people with frailty can be rapidly and accurately identified in residential care using our novel AI-derived model for frailty. A rapid identification of frailty is crucial to optimally manage the condition [27]. Indeed, the recent Australian Royal Commission to Aged Care highlighted the importance of early identification of aged care residents with frailty, who require additional support [28].

The value of any AI-derived model for frailty screening can be judged by the amount it reduces the cost of acquisition of the features required to determine the value of the deficits used to construct a frailty index. Features that are routinely collected and stored in a database in a format that can be directly fed into a classification model have a low cost of acquisition. Unfortunately, as shown in this study (Table 5) and others [20], such models lack both the sensitivity and specificity to be useful screening tests. At the other extreme, models that include all the deficit features used to calculate the EFI perform extremely well [20] (Table 4), but the value of such models is marginal.

To be useful for a screening test, a model must be acceptably accurate and significantly reduce the cost of acquisition of the features required to implement a frailty index. If a model cannot be developed with acceptable accuracy without including at least some high-cost features, it is desirable to determine the optimal minimum set of high-cost features required to achieve an acceptable performance. Genetic algorithms perform well at determining the optimal subset of features required to maximize the performance of a model. Furthermore, their choice of a subset can be limited to any number of features, up to and including all the available features. This allows the trade-off

### Table 5. Performance of models based only on low-cost features.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPA(^a)</th>
<th>NPA(^b)</th>
<th>Accuracy</th>
<th>F1(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR(^d)</td>
<td>77.3</td>
<td>63.3</td>
<td>69.1</td>
<td>72.5</td>
<td>71.1</td>
<td>66.1</td>
</tr>
<tr>
<td>SVM(^e)</td>
<td>77.3</td>
<td>58.3</td>
<td>67.3</td>
<td>69.9</td>
<td>68.9</td>
<td>62.5</td>
</tr>
<tr>
<td>DT(^f)</td>
<td>61.3</td>
<td>70.0</td>
<td>59.2</td>
<td>71.9</td>
<td>65.2</td>
<td>64.1</td>
</tr>
<tr>
<td>RF(^g)</td>
<td>77.3</td>
<td>58.3</td>
<td>67.3</td>
<td>69.9</td>
<td>68.9</td>
<td>62.5</td>
</tr>
</tbody>
</table>

\(^a\)PPA: positive percent agreement.  
\(^b\)NPA: negative percent agreement.  
\(^c\)F1: F-score.  
\(^d\)LR: logistic regression.  
\(^e\)SVM: support vector machine.  
\(^f\)DT: decision tree.  
\(^g\)RF: random forest.
predict frailty, they could be used in any domain. They are well suited to permit AI models to be trained to implement screening tests in domains where costs are important and there is a difference in the cost of acquisition of candidate features.

**Limitations**

Because this study reuses the data from a previous study [20], it shares the limitations associated with the data from the first study. In particular, the data were sourced from a single aged care provider, and the data set was relatively small. This study further filtered patients based on the availability of an ACFI assessment. It is plausible that these criteria gave a skewed representation of the population that a screening test would be applied to, resulting in different model performance. The ability to reproduce AI results continues to be controversial [29,30] within medicine, so further studies should aim to reproduce these results with different data sets. A further limitation is the changing model of aged care in Australia, with a new model set to replace ACFI in the next 2 years.

**Conclusion**

The value of screening tests lies in their cost-effective application. The main cost of applying a model-based screening test lies in acquiring the measures fed into the model. To derive useful screening tests using AI techniques, algorithms must be employed that favor the use of cheaper features over those that require more effort or patient risk to acquire. What all aged care providers and their clinical advisers need is a screening tool that will allow the efficient planning of evidence-based interventions to older frail people who will best benefit from them. At a time where the aged care sector and all providers are being asked by governments and national quality agencies to focus on this vulnerable group, it is crucial that we employ an efficient screening tool.

This paper has shown how partial genetic algorithms can be used to determine an optimal subset of high-cost features to use with cheap features to derive AI models to classify frailty, both in terms of which parameters to use and how many to use. This technique can be applied to any database. It does not guarantee that an adequate model will be found from any database, but it does give a good indication of whether there is sufficient information in the data to derive a model.

Partial genetic algorithms were demonstrated in this paper to derive a cost-effective screening test for frailty, but the method can be applied to any screening tests where there is a disparity in the cost of measuring the required features. The outcome of this study will aid health care providers in screening for frailty with better accuracy through the proposed cost-effective method, which strikes a good balance between accuracy and cost.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Full list of features.

[DOCX File . 14 KB - aging_v5j4e38464_app1.docx ]

Multimedia Appendix 2

Selected features.

[DOCX File . 41 KB - aging_v5j4e38464_app2.docx ]

Multimedia Appendix 3

Low-cost features selected for models built with GA-selected subset.

[DOCX File . 13 KB - aging_v5j4e38464_app3.docx ]

**References**


Abbreviations

ACFI: Aged Care Funding Instrument
AI: artificial intelligence
eFI: electronic frailty index
FI: frailty index
FP: frailty phenotype
GA: genetic algorithm
NLP: natural language processing
PAS: Psychogeriatric Assessment Scales
SVM: support vector machine

Please cite as:
Oates J, Shafiabady N, Ambagtsheer R, Beilby J, Seiboth C, Dent E
Evolving Hybrid Partial Genetic Algorithm Classification Model for Cost-effective Frailty Screening: Investigative Study
JMIR Aging 2022;5(4):e38464
URL: https://aging.jmir.org/2022/4/e38464
doi:10.2196/38464
PMID: