

1 **Title: Toward Predicting Peripheral Artery Disease Treatment Outcomes Using**  
2 **Non-Clinical Data**

---

3  
4 Ali Al Ramini,<sup>1</sup> Farahnaz Fallahtafti,<sup>2</sup> Mohammad Ali Takalloulraklis,<sup>3</sup> Iraklis Pipinos<sup>4,5</sup>,  
5 Sara Myers<sup>2,4</sup> & FadiAlsalem<sup>1\*</sup>

6  
7 1 Mechanical Engineering, Department, University of Nebraska-Lincoln, Lincoln, NE 68588, USA.

8 2 Department of Biomechanics, University of Nebraska at Omaha, Omaha, NE 6160, USA.

9 3 Architectural Engineering Department, University of Nebraska–Lincoln, Omaha, NE 68182, USA.

10 4 Department of Surgery and VA Research Service, VA Nebraska-Western Iowa Health Care System,  
11 Omaha, NE 68105, USA. 4

12 5 Department of Surgery, University of Nebraska Medical Center, Omaha, NE 68198, USA

13 \*email: falsaleem2@unl.edu

14 **Abstract**

---

15 Peripheral Artery Disease (PAD) significantly impairs quality of life and presents varying  
16 degrees of severity that correctly identifying would help choose the proper treatment approach  
17 and enable personalized treatment approaches. However, the challenge is that there is no  
18 single agreed-on measure to quantify the severity of a patient with PAD. This led to a trial-and-  
19 error approach to deciding the course of treatment for a given patient with PAD. This study uses  
20 non-clinical data, such as biomechanical data and advanced machine-learning techniques, to  
21 detect PAD severity levels and enhance treatment selection to overcome this challenge. Our  
22 findings in this paper lay the groundwork for a more data-driven, patient-centric approach to  
23 PAD management, optimizing treatment strategies for better patient outcomes.

24 **Keywords:** Peripheral artery disease (PAD); Ground reaction forces (GRF);  
25 Biomechanics; Walking Impairment Questionnaire (WIQ); Medical Outcomes Study  
26 Short Form 36 (SF-36); Absolute Claudication Distance (ACD); PAD severity; Machine  
27 Learning; Gait analysis

## 28 Introduction

---

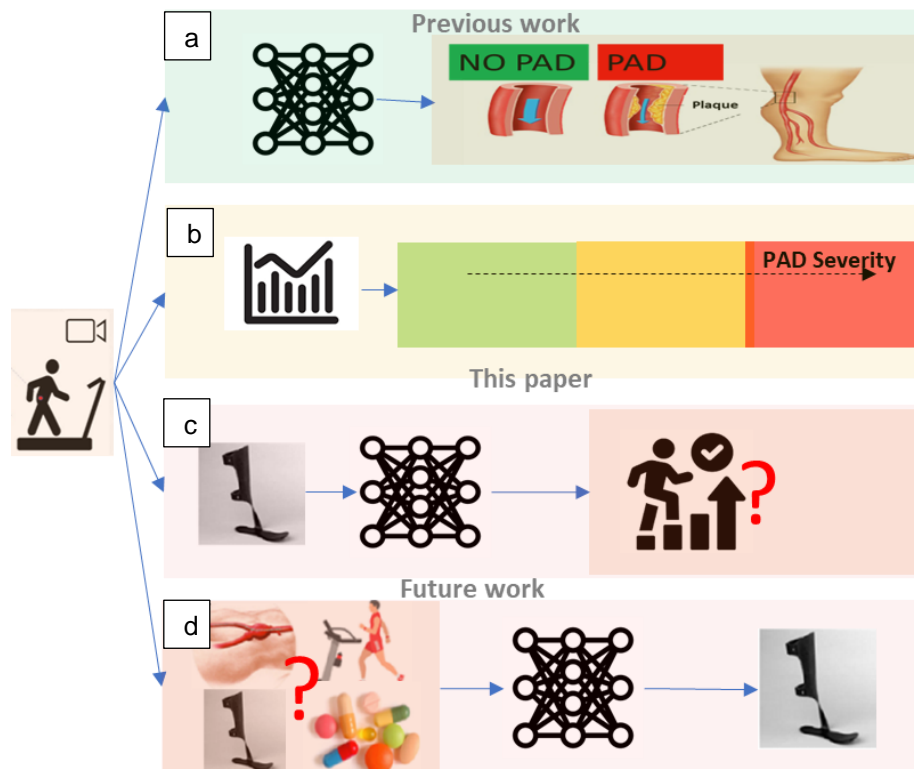
29

30 This paper presents the use of non-clinical data to characterize severity levels and  
31 support decision-making when treating chronic diseases such as peripheral artery  
32 disease (PAD). PAD is an atherosclerotic syndrome that leads to occlusion of the  
33 arteries supplying the legs. PAD affects approximately 8 million people in the US,  
34 producing a considerable public health burden.<sup>1-3</sup> The standard therapies for PAD  
35 include pharmacotherapy,<sup>4</sup> supervised exercise therapy<sup>5</sup>, wearing assistive devices<sup>6,7</sup>  
36 endovascular revascularization (angioplasty/stenting), and open revascularization  
37 (bypass operations).<sup>8</sup> Knowing which treatment will benefit each patient most is very  
38 challenging. Clinical evidence shows that treatment outcomes vary widely across  
39 patients with PAD<sup>4</sup>, and the factors contributing to the success or failure of treatments  
40 are poorly understood.<sup>1,9</sup> For example, while assistive ankle foot orthoses (AFO) show  
41 promise in enhancing walking distances, the research on its consistent use and patient  
42 perceptions is limited.<sup>10</sup> Conventional pharmacological treatments do not address  
43 existing blockages and muscle myopathy experienced by individuals with PAD and only  
44 minimally improve walking distances.<sup>11,12</sup> Revascularization lacks clear superiority  
45 between bypass surgery and endovascular revascularization, and surgical interventions,  
46 in general, have high risks to patients and varying patency rates.<sup>13-17</sup> All these  
47 limitations highlight the need for evidence-based guidelines for treating a patient with  
48 PAD.<sup>18</sup>

49 An additional challenge in PAD treatment is that there are no standard agreed-upon  
50 measures to confirm improvement after treatment. Existing outcome measures vary,

51 and the most common measures are not inclusive by only assessing blood flow  
52 improvements (ankle-brachial indices) or very subjective, such as patient questionnaires  
53 or self-reported improvement or worsening of symptoms. These limitations, when  
54 combined with PAD's complex nature, have limited physician's ability to understand  
55 which course of treatment for a given patient with PAD would lead to the best  
56 outcomes.<sup>3</sup>

57 A predictive model can significantly enhance clinicians' ability to make informed  
58 decisions, tailor interventions, and optimize post-treatment care of patients with PAD,  
59 thus improving care quality and reducing ineffective or high-risk interventions.<sup>19</sup> Toward  
60 this goal, this paper seeks to harness the power of machine learning and  
61 comprehensive gait analysis to provide a proof of concept for a data-driven approach to  
62 PAD treatment and management. Our approach leverages the data we have collected  
63 for patients with PAD over the last 20 years. The comprehensive dataset includes  
64 clinical data and gait biomechanics measurements for healthy subjects and patients  
65 with PAD before and after treatments. Our recent findings demonstrated that a machine  
66 learning approach with gait biomechanics data could accurately classify individuals as  
67 having or not having PAD (Figure 1.a).<sup>20,21</sup> Building on this finding, our current study will  
68 use comprehensive biomechanics gait data to establish reliable measures for the PAD  
69 severity level (Figure 1.b), which can then be used to develop models predicting a  
70 specific treatment outcome (Figure 1.c). These models are the first step toward building  
71 an intelligent expert system to support individualized treatment decisions for patients  
72 with PAD (Figure 1.d).



73

74 **Figure 1: Our Progress and future work toward PAD treatment prediction.** a) Our  
75 previous work showed that biomechanics data could be used to classify individuals as  
76 having or not having PAD.<sup>20,21</sup> b) This paper aims to establish a measure of PAD  
77 severity, which will be the engine to c) train models to predict treatment outcomes of  
78 patients with PAD and then in the future to d) predict the optimal treatment for patients  
79 with PAD.

80

## 81 Results

82 Figure 2. a presents a comprehensive distribution of all features in our study for healthy  
83 controlled and patients with PAD before different treatments (AFO and Surgery). The  
84 features are represented after the post-application of the Yeo-Johnson transformation. It  
85 is evident from the figure that kinematics features (ankle, hip, and knee) failed to  
86 differentiate between healthy individuals and those with PAD. While qualitative  
87 questionnaire evaluations, such as SF-36 and WIQ, effectively distinguished between

88 healthy controls and patients with PAD, they failed to discern PAD severity, particularly  
89 when differentiating between AFO and surgery patients. The severity (supported by  
90 clinical measures such as the ABI test) is ranked based on the type of intervention  
91 performed on a patient with PAD: Healthy - AFO - Surgery. In other words, the fact that  
92 a patient with PAD has undergone open surgery means a vascular physician  
93 determined his PAD severity level to be higher than the patient who had just AFO. The  
94 GRF features distinguished all three categories and presented an ordered patient  
95 distribution. This observation highlights the potential of GRF features as a potent  
96 instrument for quantifying PAD severity.

97 Figure 2.b presents the outcomes of the Mann-Whitney U test aiming to recognize  
98 statistically significant differences among the Healthy, AFO, and Surgery groups. The  
99 results confirm the observations (Figure 2.a) that GRF features emerge as most able to  
100 distinguish across groups, exhibiting the most significant statistical differences between  
101 the three groups in the pre-intervention phase. In contrast, hip, ankle, and knee  
102 kinematic features do not consistently capture the disparities between these groups.  
103 Furthermore, the WIQ questionnaire features effectively differentiate between the  
104 Healthy group and patients with PAD groups, but not between the AFO and Surgery  
105 groups.

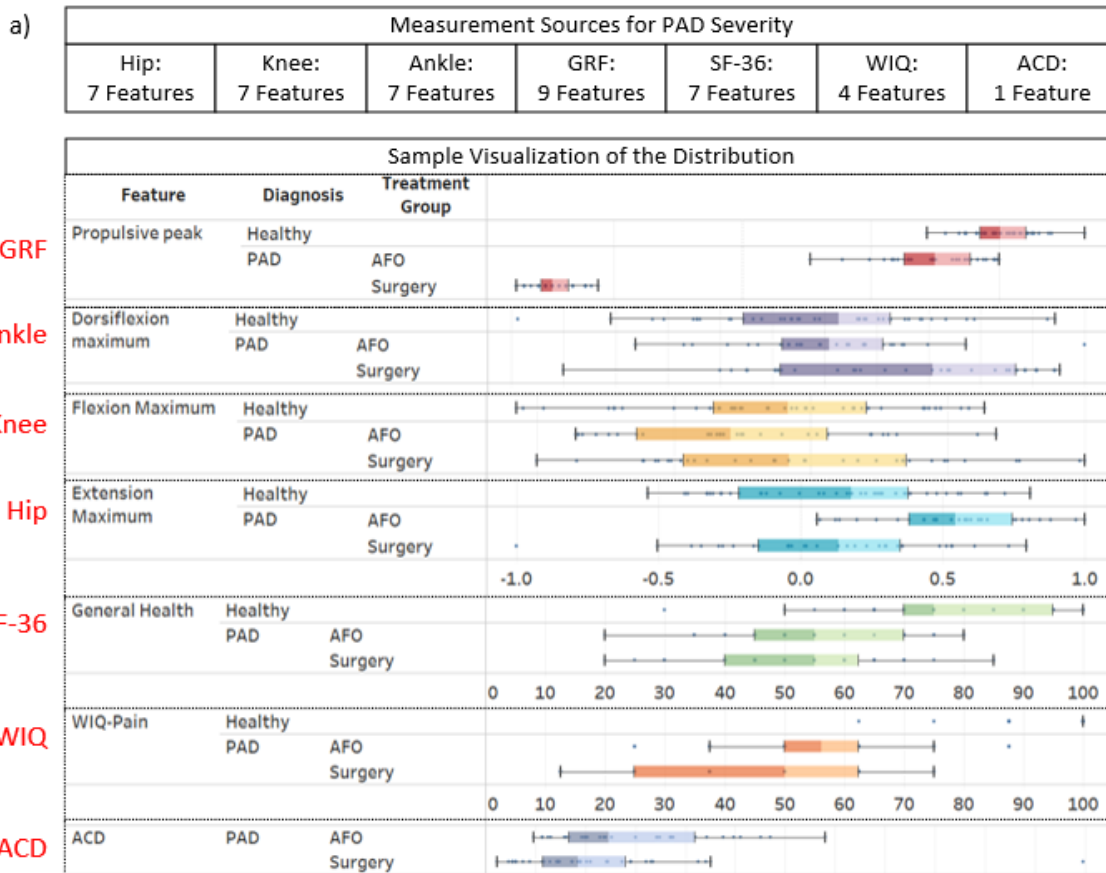
106 Next, the results of using machine learning modes to decode PAD severity using the  
107 different features are shown in (Figure 2.c). The GRF, using logistic regression with a  
108 single feature, "Propulsive peak," achieved an accuracy<sup>22</sup> of 0.909, a balanced  
109 accuracy<sup>22</sup> of 0.867, and a Matthews Correlation Coefficient (MCC)<sup>22</sup> of 0.868. The  
110 GRF's high performance across these metrics, particularly the MCC that accounts for

111 unbalanced data, highlights its effectiveness in accurately quantifying PAD severity.  
112 Additionally, the consistent results between the model's training and testing  
113 performance suggest that the GRF situation effectively captures the essential patterns  
114 in the data without being overly fitted to the training set or missing the situation's  
115 complexity. This confirms the robustness of GRF features for reliable severity  
116 assessment in PAD.

117 Figure 2.d aims to provide a more granular understanding of PAD severity by leveraging  
118 the GRF Propulsive Peak, the best-performing feature identified in the previous  
119 analysis, by stratifying a PAD scale. The data analysis yielded a discernible stratification  
120 of PAD severity based on GRF Propulsive Peak values (Figure 2.d) derived from the  
121 logistic regression model predictions. The resulting figure defines three distinct regions:

- 122 • Surgery Region (More Severe PAD): Ranging from -1 to 0.075.
- 123 • AFO Region (Less Severe PAD): Extending from 0.075 to 0.775.
- 124 • Healthy Region: Above 0.775 on the GRF Propulsive Peak scale.

125  
126 This stratification visually represents how GRF Propulsive Peak values correlate with  
127 varying degrees of PAD severity. In summary, the results in Figure 2 show the  
128 importance of GRF as a robust and reliable data source for PAD severity quantification,  
129 outperforming other commonly used metrics and questionnaires such as SF-36 and  
130 WIQ.



131

b) Statistical Testing: Comparison with Control Group

Measurement Source	Number of Features	Pair	
GRF	9	Healthy vs pre AFO	0.003
		Healthy vs pre Surgery	0.000
		AFO vs pre Surgery	0.000
Ankle	7	Healthy vs pre AFO	0.018
		Healthy vs pre Surgery	1.000
		AFO vs pre Surgery	0.559
Hip	7	Healthy vs pre AFO	0.000
		Healthy vs pre Surgery	0.206
		AFO vs pre Surgery	0.000
Knee	7	Healthy vs pre AFO	1.000
		Healthy vs pre Surgery	0.075
		AFO vs pre Surgery	1.000
WIQ	4	Healthy vs pre AFO	0.000
		Healthy vs pre Surgery	0.000
		AFO vs pre Surgery	1.000
SF-36	7	Healthy vs pre AFO	0.000
		Healthy vs pre Surgery	0.000
		AFO vs pre Surgery	1.000
ACD	1	AFO vs pre Surgery	0.030

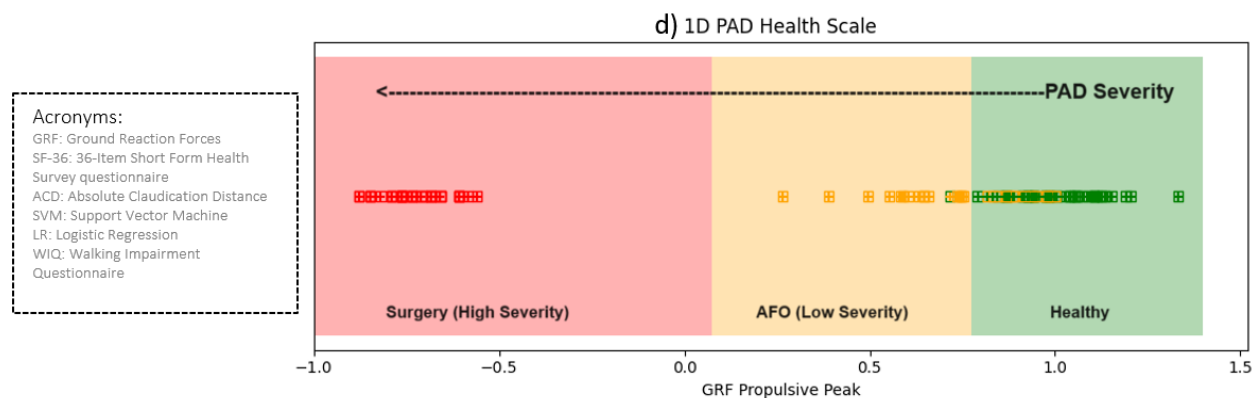
c) Employing Machine Learning to Decode PAD Severity

Feature	Number of Features	Best Model	Accuracy	MCC
GRF-All	9	SVM	0.86	0.79
Braking peak	1	SVM	0.86	0.79
Propulsive peak	1	LR	0.91	0.87
Propulsive impulse	1	LR	0.86	0.79
Braking impulse	1	SVM	0.77	0.65
Lateral peak	1	SVM	0.77	0.66
Medial peak	1	LR	0.77	0.69
Loading response	1	LR	0.77	0.69
Midstance valley	1	LR	0.77	0.69
Terminal stance peak	1	LR	0.77	0.69
SF-36-All	7	LR	0.59	0.40
Emotional Well Being	1	LR	0.46	0.12
General Health	1	LR	0.64	0.44
Emotional Problems	1	SVM	0.50	0.23
Physical Health	1	SVM	0.59	0.36
Pain	1	LR	0.82	0.73
Physical Function	1	LR	0.59	0.40
Social Functioning	1	LR	0.64	0.44
WIQ-All	4	SVM	0.73	0.58
WIQ-Distance	1	LR	0.82	0.74
WIQ-Pain	1	SVM	0.82	0.77
WIQ-Speed	1	LR	0.77	0.68
WIQ-Stair Climbing	1	SVM	0.64	0.44



132





133

134 **Figure 2: GRF emerges as a strong indicator of PAD severity.** a) Pre-intervention  
135 distribution of all features categorized based on diagnosis and treatment groups  
136 (Healthy, AFO, and Surgery). b) Results of the Mann-Whitney U test comparing the  
137 Healthy, AFO, and Surgery groups across various gait features. The figure showcases  
138 each measurement source's Median Bonferroni P Value, with lower values indicating  
139 higher statistical significance. c) Comparative performance of machine learning models  
140 in feature-level PAD severity quantification across all features as independent inputs. d)  
141 Stratification of PAD Severity Based on GRF Propulsive Peak Values: A continuous  
142 scale derived from Logistic Regression model predictions, delineating three distinct  
143 regions - Healthy (-1 to -0.8), AFO (Less Severe PAD, -0.8 to 0.25), and Surgery (More  
144 Severe PAD, 0.25 to 1). This scale correlates GRF Propulsive Peak values and PAD  
145 severity, facilitating swift and precise clinical assessments.

146

147 Next, interest is shifted to understanding the patients with PAD's response to  
148 interventions using effect size analysis<sup>23</sup>. This analysis quantitatively represents the  
149 strength of the relationship between variables using Cohen's d measure. Cohen's  
150 guidelines suggest that a d value of 0.2 indicates a 'small' effect size, 0.5 a 'medium'  
151 effect size, and 0.8 a 'large' effect size. In our context, with a control group and two  
152 patient groups (Surgery and AFO), both pre and post-intervention, the effect size aids in  
153 quantifying the magnitude of change due to intervention in each patient group and offers  
154 a comparative measure against the control. For instance, comparisons such as Healthy  
155 vs. pre-surgery provide insights into the deviation of the surgery group from the control



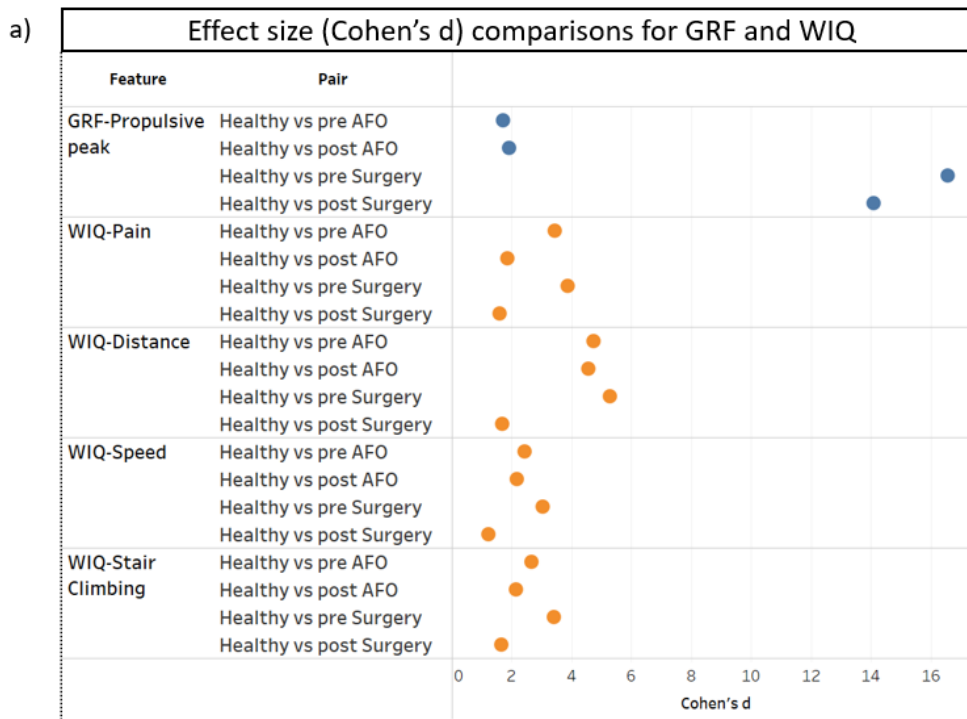
156 before the intervention. At the same time, Healthy vs. post-surgery reveals the deviation  
157 post-intervention.

158 Figure 3.a presents the effect sizes, quantified using Cohen's d, for one GRF measure  
159 and four WIQ measures as an example relative to the healthy control group. Figure 3.a  
160 shows the magnitude of this effect size based on GRF Propulsive Peak, confirming the  
161 severity level differences between patients with PAD who had surgery versus patients  
162 completing an AFO intervention. More importantly, the effect size decreases post-  
163 surgery, suggesting a shift towards the healthy group's values. In contrast, the AFO  
164 group shows minimal change in the GRF values post-intervention. Both behaviors are  
165 somewhat expected, extending the confirmation that GRF can also be used to capture  
166 PAD status after intervention. However, the WIQ fails to capture the anticipated severity  
167 levels among patients with PAD following AFO and surgery. In addition, their effect size  
168 indices project potentially over-optimistic results, with post-surgical and post-AFO  
169 cohorts yielding equivalent Cohen's d magnitudes.

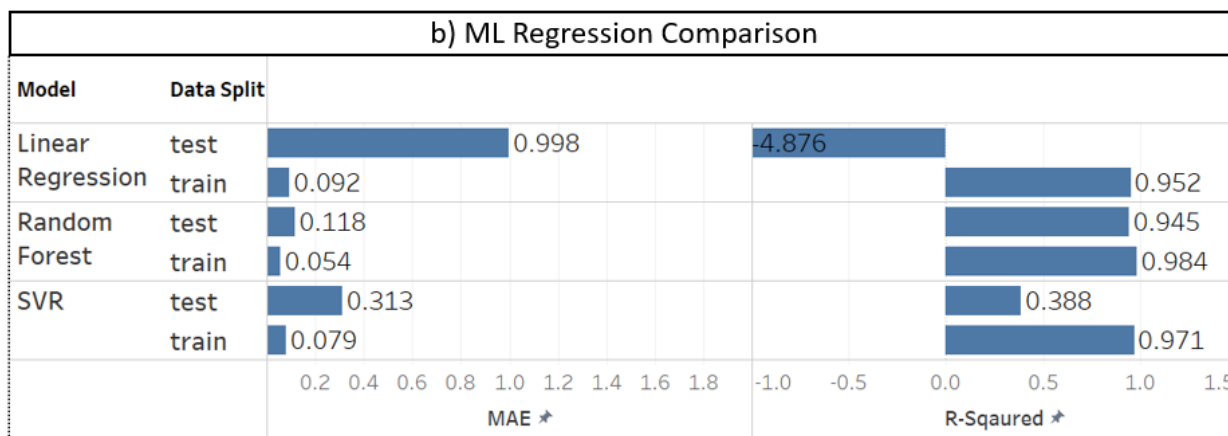
170 Next, the most profound part of this paper is that machine learning models were trained  
171 using pre-intervention data, including gait kinematics and kinetics measurements, WIQ  
172 scores to predict post-intervention effectiveness. Post-intervention effectiveness is  
173 captured by improvement in GRF Propulsive Peak. Figure 3b results provide a  
174 comprehensive overview of the performance metrics for various machine-learning  
175 models (Linear Regression, Random Forest, and SVM). Figure 3.c shows that the  
176 Random Forest model consistently outperforms the other models in Mean Absolute  
177 Error (MAE) and R-squared. Furthermore, we added a binary representation for the  
178 model outcomes by assessing if the predicted values can specify whether a patient's

179 situation improved based on the predicted and actual data results. The confusion  
180 matrices for both training and test datasets, as depicted in Figure 3.c, provide a clear  
181 visual representation of the model's predictions against the actual outcomes. For the  
182 training data, the model achieves an accuracy of 86%, while for the test data, the  
183 accuracy stands at 82%. These high accuracy scores indicate the model's ability to  
184 assess treatment outcomes' direction. This indicates that machine-learning applications  
185 could be practical in predicting treatment outcomes for PAD.

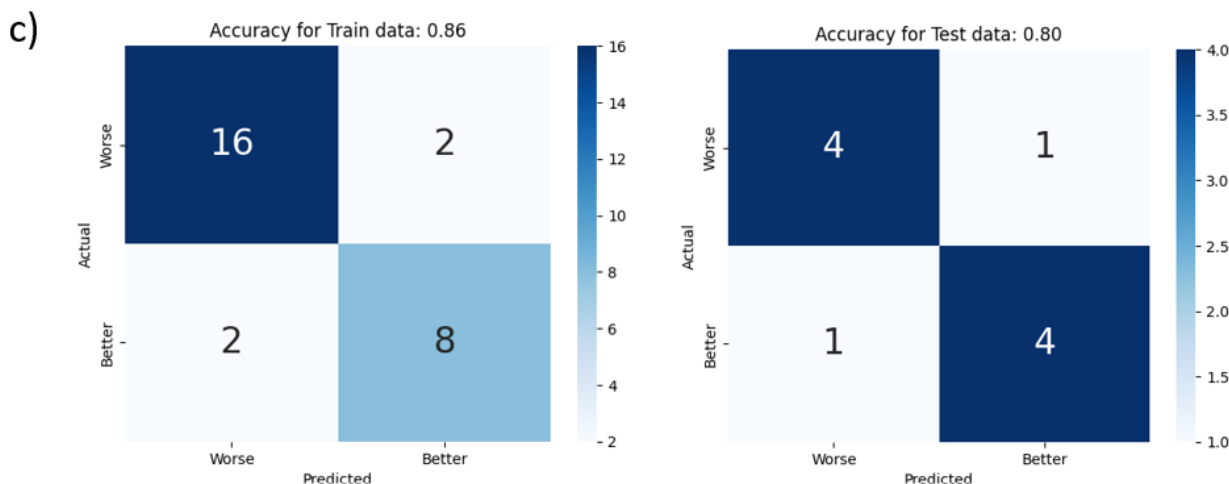
186 The individual predictions of GRF Propulsive Peak post-intervention for each subject, as  
187 shown in Figure 3.d, offer a granular view of the model's performance. Each subject's  
188 data point is plotted against their pre-intervention GRF Propulsive Peak, with distinct  
189 markers indicating the actual post-intervention values and the model's predictions. A  
190 consistent alignment between the actual and predicted values is evident across most  
191 subjects. This individual-level analysis complements the accuracy metrics, reinforcing  
192 the model's ability to predict treatment outcomes. The colored bars visually represent  
193 the model's confidence in its predictions, with the proximity of the actual and predicted  
194 markers indicating the precision of the model's forecasts.



195

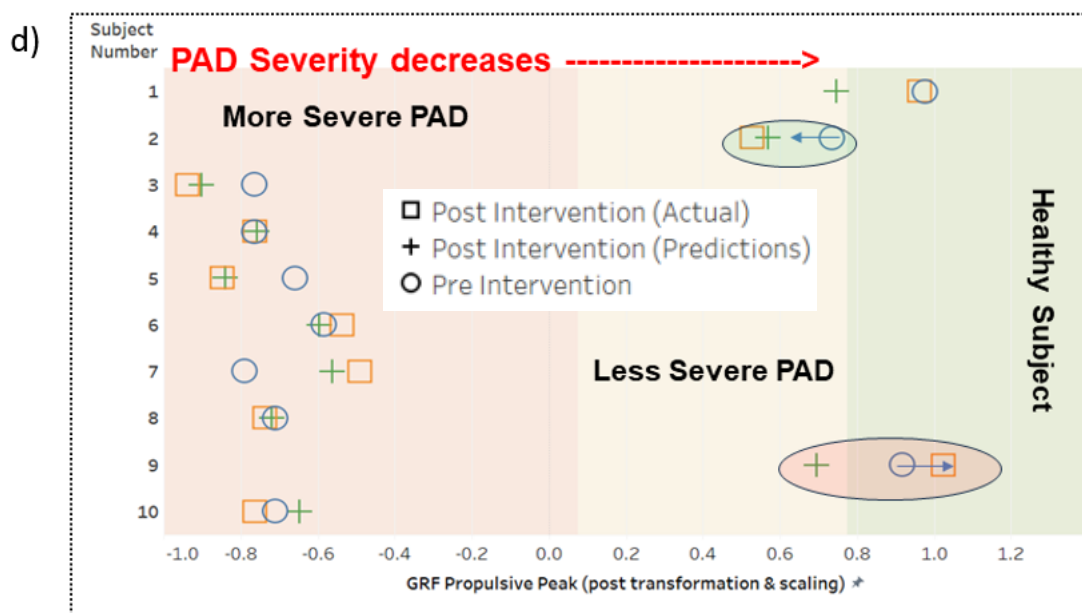


196



197

198



199 **Figure 3: Machine learning models are developed to predict the outcome of a**  
200 **certain treatment.** a) Effect size (Cohen's d) comparisons for GRF, WIQ, and SF-36  
201 scores between healthy individuals and patients with PAD, both pre and post-  
202 intervention. Higher Cohen's d values indicate larger differences between the groups,  
203 suggesting more pronounced effects of interventions or more distinct group  
204 characteristics. Blue dots represent comparisons involving the 'pre' status, and orange  
205 dots signify the 'post' status. b) Performance metrics comparison of machine learning  
206 models including Linear Regression, Random Forest, and SVR. Metrics shown are  
207 MAE, R-Squared, and adjusted R-Squared, assessing model accuracy and predictive  
208 capability with both Correlation and PCA feature selection methods on training and test  
209 datasets. c) Confusion matrices for treatment outcome predictions (Worse vs. Better)  
210 using the Random Forest model on training and test datasets, showing the counts of  
211 correct and incorrect predictions. d) Individual Predictions of GRF Propulsive Peak for  
212 All Subjects Post-Intervention. The graph plots the actual vs. predicted GRF Propulsive  
213 Peak values, illustrating prediction accuracy. The spectrum of 'More Severe PAD' to  
214 'Less Severe PAD' visualizes the range of severity based on GRF values, with a clear  
215 demarcation showing patients transitioning towards a 'Healthy' status. Notably, the  
216 distribution of cases in 'More Severe' and 'Less Severe' PAD is not purely based on the  
217 number of surgeries or AFO interventions but on the observed severity metrics in the  
218 GRF Propulsive Peak values post-treatment.

## 219 Discussion

220 The inability of simple tests to differentiate between healthy controls and patients with  
221 PAD emphasizes the need for more robust metrics in estimating a patient with PAD

222 severity level. While qualitative assessments like SF-36 and WIQ have their merits, their  
223 inability to discern between AFO and Surgery patients raises questions about their  
224 efficacy. Remarkably, the GRF Propulsive Peak feature stands out, with the ability to  
225 differentiate across all categories and an ordered distribution among patients with PAD  
226 that suggests their potential as a foundation of characterizing PAD severity. The GRF  
227 Propulsive Peak introduces a straightforward 1D scale for measuring PAD severity. This  
228 1D scale is precious for its practical applications outside clinical settings, offering a  
229 simple and effective tool for monitoring PAD severity. The 1D scale is a beacon for  
230 future innovations that aim to make PAD management more efficient and patient-  
231 centric.

232 GRF Propulsive peak values using the effect size test show a better correlation with  
233 post-treatment interventions. On the other hand, qualitative assessments such as WIQ  
234 scores might present an overly optimistic view of the patient's condition. This asserts  
235 the importance of integrating GRF measurement with subjective patient-reported  
236 outcomes to understand PAD severity and its response to interventions.

237 The Random forest model accurately predicts the post-intervention GRF propulsive  
238 peak. Using such a model, clinicians can gain valuable insights into the potential  
239 success of a treatment before its actual implementation. This predictive capability can  
240 revolutionize patient care, allowing for more personalized treatment plans and  
241 potentially reducing the number of ineffective interventions. Moreover, it is a  
242 foundational step towards a more data-driven approach in PAD treatment, where  
243 decisions are informed by predictive analytics rather than merely relying on traditional  
244 methods. As we continue to refine and validate this model, it paves the way for more

245 comprehensive studies that can further unravel the complexities of PAD and optimize  
246 treatment strategies for better patient outcomes.

247 The ability to predict individual treatment outcomes with high accuracy, as  
248 demonstrated by the Random Forest model, holds significant implications for clinical  
249 practice. Clinicians can effectively tailor their treatment approaches and post-treatment  
250 care by estimating the post-intervention GRF Propulsive Peak. The model's  
251 performance, both in terms of overall accuracy and individual predictions, suggests that  
252 it can be a valuable tool in decision-making. As we delve deeper into personalized  
253 medicine, such predictive capabilities become increasingly crucial. The alignment  
254 between the model's predictions and the actual outcomes underscores the potential of  
255 integrating machine learning into PAD treatment strategies. This integration enhances  
256 the precision of treatment planning and facilitates more informed patient-clinician  
257 discussions, fostering a collaborative approach to care.

258 While providing valuable insights, our study has certain limitations. Firstly, the dataset  
259 employed is relatively constrained, encompassing only 97 subjects, of which 42 are  
260 healthy controls, and 65 are patients diagnosed with PAD. Consequently, our analytical  
261 and machine learning outcomes necessitate validation through a broader dataset with a  
262 greater patient count and severity variability. Secondly, the result encapsulated in the  
263 GRF Propulsive peak scale offers room for refinement. The current scale predominantly  
264 categorizes patients with PAD into two severity brackets: less severe (AFO patients)  
265 and more severe (Surgery patients). A more nuanced representation could be achieved  
266 by incorporating a broader spectrum of PAD severity gradations.

267 Future research can add more patients with PAD data from other treatments to the  
268 model to improve generalization and provide more options for predicting treatment  
269 outcomes. Moreover, real-time GRF data could provide a more nuanced understanding  
270 of PAD severity and post-treatment effect.

271

## 272 Methods

---

### 273 **Data Sources and Description**

274 Biomechanics<sup>24</sup> data for this study were sourced from research approved by the  
275 Institutional Review Boards at the University of Nebraska Medical Center and the  
276 Nebraska-Western Iowa Veteran Affairs Medical Center, involving 65 individuals with  
277 PAD and 42 healthy controls. Of the patients with PAD, 9 were treated with AFO, and  
278 35 underwent surgical treatments. Some AFO-treated patients' data were collected  
279 multiple times pre-intervention. Therefore, we have 30 records for AFO-treated patients  
280 before intervention. Not all patient data were collected after the intervention; 30 surgery  
281 patients' data were collected after the surgery, and 8 AFO-treated patients' data were  
282 collected after three months of the treatment. In this study, we considered patients  
283 treated with AFOs to have less severe PAD than those treated with surgery.

284 In addition to biomechanical data, qualitative and quantitative assessments were  
285 conducted to understand the PAD condition better. The WIQ was administered to gauge  
286 the self-perceived walking ability of patients with PAD<sup>25-27</sup>. This questionnaire evaluates  
287 various aspects of walking, including pain, distance, speed, and stair climbing, providing  
288 insights into patients' daily challenges with PAD. SF-36 was also employed to assess



289 the participant's health status and quality of life<sup>12,28</sup>. This comprehensive questionnaire  
290 evaluates physical and mental health domains, offering a broader perspective on the  
291 impact of PAD on patients' daily lives.

292 Furthermore, the Absolute Claudication Distance<sup>26</sup> test was conducted. This quantitative  
293 assessment measures the distance a patient can walk before being compelled to stop  
294 due to claudication pain, providing a direct metric of the severity of PAD symptoms.  
295 Collectively, these assessments, combined with the biomechanical data, aimed to offer  
296 a multi-dimensional perspective on PAD's impact on patients. Furthermore, this multi-  
297 dimensional dataset offers an opportunity to compare various measurement sources to  
298 identify the most precise metric for assessing PAD severity.

299 For this analysis, multiple measurement sources were considered:

300 1. Ankle, Hip, and Knee Kinematics: Each anatomical region was characterized by  
301 seven distinct features, capturing the nuances of movement and biomechanical  
302 alterations.<sup>29</sup>

303 2. Ground Reaction Forces (GRF): Nine features were extracted to understand the  
304 forces exerted during walking, providing insights into gait alterations.<sup>24</sup>

305 3. Walking Impairment Questionnaire (WIQ): Four features were derived from this  
306 self-administered questionnaire, offering a patient-centric perspective on walking  
307 ability.<sup>26</sup>

308 4. Medical Outcomes Study Short Form 36 (SF-36): Seven features were  
309 considered from this tool, gauging the participants' health status and quality of life.

310 5. Absolute Claudication Distance (ACD): This singular feature was exclusively  
311 available for patients with PAD, measuring the distance they could walk before the  
312 claudication pain.

### 313 **Data Transformation**

314 In the preliminary stages of our analysis, we identified the need to standardize and  
315 transform our biomechanical data to ensure comparability across various biomechanics  
316 features and conditions. We employed the Yeo-Johnson transformation<sup>30</sup>, a method  
317 optimized for varying data distributions, including zero and negative values. Initially, this  
318 transformation was determined and performed based on the pre-intervention dataset.  
319 Subsequently, the same transformation parameters derived from the pre-intervention  
320 data were applied to both the post-intervention and healthy datasets, ensuring  
321 consistency across all conditions. After the transformation, we implemented feature  
322 scaling to standardize the data distribution further, which is crucial for the performance  
323 of certain machine learning models and the validity of statistical tests that assume data  
324 uniformity. This standardization facilitates more accurate analyses and helps in  
325 achieving reliable results. This scaling was configured using the transformed pre-  
326 intervention data, ensuring that all biomechanical features ranged between -1 and 1.  
327 The same scaling parameters were then applied to the post-intervention and healthy  
328 data. This systematic data transformation and scaling approach enhanced data clarity,  
329 separation, and comparability across all conditions and features.

### 330 **Statistical Comparison with Control Group Using the Mann Whitney U-test.**

331 We employ the non-parametric Mann-Whitney U-test<sup>31</sup> to distinguish the differences in  
332 measurements between the three independent groups - Healthy, AFO, and Surgery.

333 The Healthy group comprised individuals without PAD, serving as the control, while the  
334 AFO (low severity) and surgery (high severity) groups represented patients with PAD  
335 varying in degrees of disease severity.

336 We chose The Mann-Whitney U-test due to its robustness in comparing non-normally  
337 distributed data, which is often the case in medical research. In addition, many features  
338 in our data have shown non-normality. We present the results from this test as the  
339 Median Bonferroni P-value for each measurement source, which adjusts the median of  
340 the observed p-values to account for multiple comparisons, reducing the chances of  
341 type I errors in our statistical analysis.<sup>32</sup> It is worth noting that the ACD data was only  
342 available for patients with PAD. Thus, comparisons involving the healthy group did not  
343 apply to this measure.

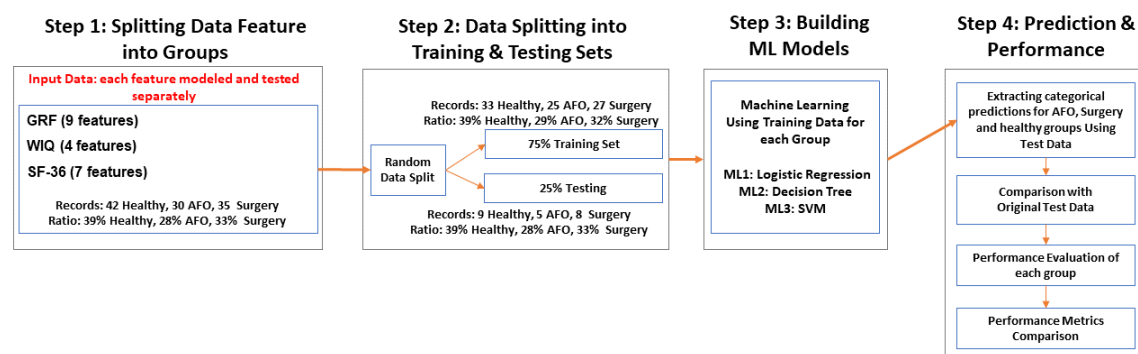
344 This statistical approach aims to identify the measurement source that best  
345 differentiates between the three groups. Ideally, the most informative measurement  
346 source would be the one that exhibits significant differences across all three pairwise  
347 comparisons: Healthy vs. AFO, Healthy vs. Surgery, and AFO vs. Surgery. For  
348 instance, if the GRF features consistently show significant disparities among these three  
349 comparisons, it would suggest that GRF features are the most potent identifiers of PAD  
350 severity.

### 351 **Machine learning approach to evaluate features' ability to quantify PAD severity**

352 In this section, we take a more granular approach to PAD severity classification by  
353 focusing on individual features within each data source—GRF, WIQ, and SF-36. Each  
354 feature is modeled and tested separately. This allows us to isolate the predictive power

355 of each feature, providing a more subtle understanding of its role in PAD severity  
356 classification.

357 The data is initially segmented into individual features and then subjected to a machine  
358 learning pipeline involving data splitting, data preprocessing, hyperparameter tuning,  
359 model training, and performance evaluation. We continue to employ Logistic  
360 Regression<sup>33,34</sup>, Decision Tree<sup>35</sup>, and Support Vector Machine (SVM)<sup>36</sup> models for this  
361 analysis. The goal is to identify the most informative individual features for PAD severity  
362 quantification, offering insights into the potential for streamlined, feature-centric  
363 diagnostic approaches. This feature-level assessment aims to refine our interpretation  
364 of which aspects of GRF, WIQ, and SF-36 most relate to PAD severity. It could thus be  
365 a focal point in future diagnostic and severity assessment tools.



366  
367 **Figure 4: Flowchart overview of feature-level assessment for PAD severity**  
368 **quantification.** The chart compares individual features using logistic regression,  
369 decision tree, and SVM models.

### 370 Stratification of PAD Severity Using Optimal Features

371 This section aims to provide a more granular understanding of PAD severity by  
372 leveraging the best-performing feature identified in the previous analysis. This 1D scale  
373 could be a valuable reference for future engineering applications to monitor PAD

374 severity and progression outside of the clinical environment, mainly because this feature  
375 is derived from GRF that can be physically measured in a lab setting. We stratify a 1D  
376 PAD scale based on the Logistic Regression model built using the GRF Propulsive  
377 Peak feature described in Figure 2.c. We synthesized a hypothetical dataset to  
378 elucidate the relationship between “GRF Propulsive Peak” values and PAD severity.  
379 This hypothetical dataset spanned ‘GRF Propulsive Peak’ values from -1 to 1.4,  
380 incremented by 0.025. Leveraging our optimal Logistic Regression model, previously  
381 identified as the most potent single feature model for this task, we processed this  
382 dataset to generate PAD severity predictions. We categorized these predictions into our  
383 multiclass classification groups: Healthy, AFO, and Surgery. The objective was to  
384 establish a continuous scale that could intuitively represent PAD severity based on GRF  
385 Propulsive Peak values.

### 386 **Quantifying Intervention Impact Using Effect Size**

387 Effect size<sup>23</sup> is a quantitative representation of the strength of the relationship between  
388 variables. Unlike statistical tests that merely confirm the existence of an effect or  
389 relationship, the effect size elucidates its magnitude, independent of sample size. In our  
390 context, with a control group and two patient groups (Surgery and AFO), both pre and  
391 post-intervention, the effect size aids in quantifying the magnitude of change due to  
392 intervention in each patient group and offers a comparative measure against the control.  
393 We employ Cohen’s *d*, a measure calculated as the difference between two means  
394 divided by the pooled standard deviation to quantify the differences between our  
395 groups. For instance, comparisons such as Healthy vs. pre-surgery provide insights into  
396 the deviation of the surgery group from the control before the intervention. At the same

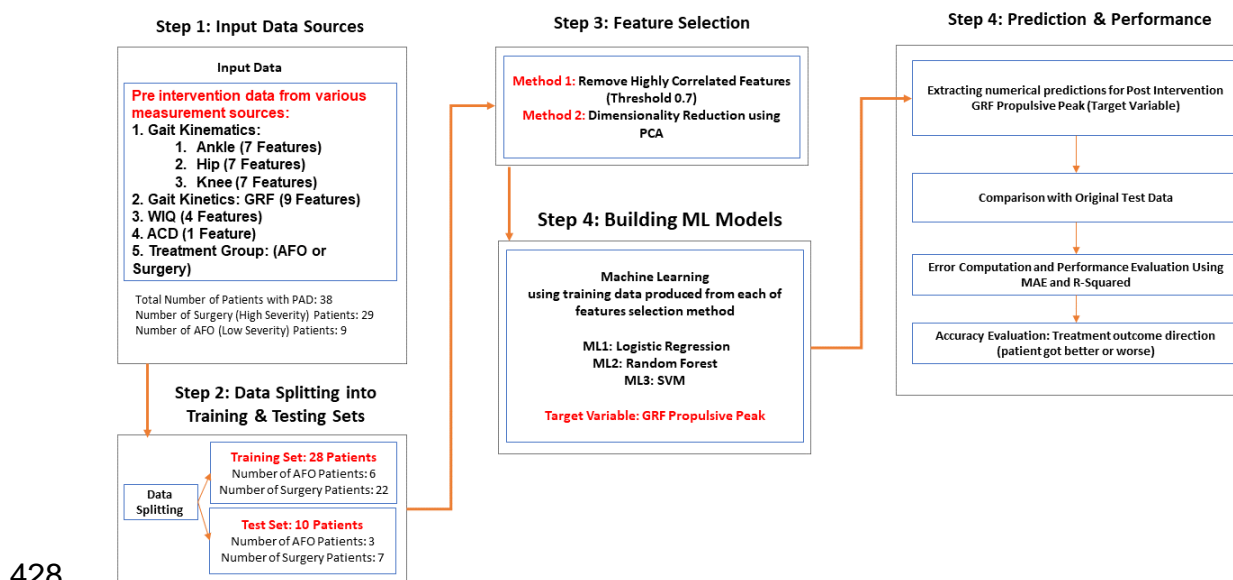
397 time, Healthy vs. post-surgery reveals the deviation post-intervention. Analogously,  
398 comparisons for the AFO group, Healthy vs. pre-AFO and Healthy vs. post-AFO provide  
399 similar insights. Cohen's guidelines suggest that a  $d$  value of 0.2 indicates a 'small'  
400 effect size, 0.5 a 'medium' effect size, and 0.8 a 'large' effect size. The sign of Cohen's  
401  $d$  further provides directional information: a positive sign indicates the post-intervention  
402 group has higher scores than the pre-intervention or control group. In contrast, a  
403 negative sign indicates the opposite. This approach aims to provide a nuanced,  
404 quantitative understanding of PAD severity and its response to interventions.

#### 405 **Machine Learning Approach Estimating Post-Treatment Outcomes**

406 The section utilizes pre-intervention data, including gait kinematics and kinetics  
407 measurements, WIQ scores, and ACD as predictive features. The initial step involves  
408 gathering pre-intervention data from a cohort of 38 patients diagnosed with PAD. This  
409 data is sourced from various measurements, including gait kinematics covering ankle,  
410 hip, and knee features and gait kinetics focusing on ground reaction forces (GRF), WIQ,  
411 and ACD. Additionally, patients are categorized based on their treatment group: AFO for  
412 those with low severity and surgery for those with high severity. Out of the total, 30  
413 patients underwent surgery due to their high severity, while eight were treated with AFO  
414 due to their lower severity.

415 The data is split into training and testing sets. The training set comprises 28 patients,  
416 with six from the AFO group and 22 from the surgery group. The remaining ten patients,  
417 consisting of 3 AFO and seven surgery patients, are reserved for the test set. This  
418 division allows the model to be trained on a diverse data set and subsequently validated  
419 on unseen data to gauge its predictive accuracy.

420 Before feeding the data into machine learning models, it undergoes a feature selection  
421 process to enhance its predictive capability. The method employs Principal Component  
422 Analysis (PCA)<sup>37</sup> for dimensionality reduction. PCA transforms the original features into  
423 a set of linearly uncorrelated variables, capturing the most significant patterns in the  
424 data while reducing its complexity (90% of the total variance is maintained). We applied  
425 this feature selection method at the training data source level to ensure that every data  
426 source contributed to the study while avoiding data leakage between the train and test  
427 sets.



429 **Figure 5: Machine learning flowchart.** The chart depicts the machine learning  
430 approach for predicting PAD treatment outcomes using pre-intervention data and  
431 evaluating model performance.

432 Three machine learning models were trained on the features of Linear Regression,  
433 Support Vector Regression (SVR), and Random Forest<sup>38</sup>. These models were chosen  
434 for their versatility and capability to handle complex nonlinear relationships in the data.  
435 The trained models were then used to predict the GRF Propulsive Peak post-



436 intervention, which, as established in previous chapters, serves as a measure of PAD  
437 severity and a reference for treatment outcome.

438 Three machine learning models are trained with optimized features: Logistic  
439 Regression, Random Forest, and Support Vector Machine (SVM). Each model is  
440 trained using the data produced from both feature selection methods. Post-training, the  
441 models predict the post-intervention GRF Propulsive Peak (established in the previous  
442 chapter, which is the most distinctive measure for PAD severity), serving as the target  
443 variable. These predictions are then compared with the original test data. The model's  
444 performance is evaluated using MAE for average prediction error magnitude and R-  
445 squared for the proportion of variance explained by the model. Additionally, the  
446 accuracy of the treatment outcome direction, indicating whether a patient's condition  
447 improves or worsens, is assessed.

448

## 449 [References](#)

- 
- 450 1. Soyoye, D. O. *et al.* Diabetes and peripheral artery disease : A review. **12**, 827–  
451 838 (2021).
- 452 2. Lin, J., Chen, Y., Jiang, N., Li, Z. & Xu, S. Burden of Peripheral Artery Disease  
453 and Its Attributable Risk Factors in 204 Countries and Territories From 1990 to  
454 2019. *Front. Cardiovasc. Med.* **9**, (2022).
- 455 3. Golledge, J. Update on the pathophysiology and medical treatment of peripheral  
456 artery disease. *Nat. Rev. Cardiol.* **19**, 456–474 (2022).
- 457 4. Huisinga, J. M., Pipinos, I. I., Stergiou, N. & Johanning, J. M. Treatment with

- 458 pharmacological agents in peripheral arterial disease patients does not result in  
459 biomechanical gait changes. *J. Appl. Biomech.* **26**, 341–348 (2010).
- 460 5. McDermott, M. M. *et al.* Six-Minute Walk Is a Better Outcome Measure Than  
461 Treadmill Walking Tests in Therapeutic Trials of Patients With Peripheral Artery  
462 Disease. *Circulation* **130**, 61–68 (2014).
- 463 6. Dinkel, D. *et al.* Assessing wear time and perceptions of wearing an ankle foot  
464 orthosis in patients with peripheral artery disease. *PM&R* **15**, 493–500 (2023).
- 465 7. Myers, S. *et al.* Examining Ankle Foot Orthosis Wear Time in Patients With  
466 Peripheral Artery Disease. *Innov. Aging* **4**, 211–211 (2020).
- 467 8. Scatena, A. *et al.* Bypass surgery versus endovascular revascularization for  
468 occlusive infrainguinal peripheral artery disease: a meta-analysis of randomized  
469 controlled trials for the development of the Italian Guidelines for the treatment of  
470 diabetic foot syndrome. *Acta Diabetol.* (2023) doi:10.1007/s00592-023-02185-x.
- 471 9. Rontoyanni, V. G. *et al.* Mitochondrial Bioenergetics in the Metabolic Myopathy  
472 Accompanying Peripheral Artery Disease. *Front. Physiol.* **8**, (2017).
- 473 10. Klukowska, A. M., Schröder, M. L., Stienen, M. N. & Staartjes, V. E. Objective  
474 functional impairment in lumbar degenerative disease: concurrent validity of the  
475 baseline severity stratification for the five-repetition sit-to-stand test. *J. Neurosurg.*  
476 *Spine* **33**, 4–11 (2020).
- 477 11. Yentes JM, Huisinga JM, Myers SA, Pipinos II, Johanning JM, Stergiou N.  
478 Pharmacological treatment of intermittent claudication does not have a significant  
479 effect on gait impairments during claudication pain. *J Appl Biomech.*

- 480 2012;28(2):184-191. doi:10.1123/jab.28.2.184
- 481 12. Ware Jr, J. E., Kosinski, M. & Gandek, B. *The SF-36 Health Survey: Manual and*  
482 *Interpretation Guide*. (Quality Metric Inc., 2000).
- 483 13. A, Eghbalieh SDD, Dardik A. Basic data related to surgical infrainguinal Contact  
484 PD/PI: Rahman, Hafizur References Cited Page 8 revascularization procedures:  
485 A twenty year update. *Ann Vasc Surg*. 2011;25(3):413-422.  
486 doi:10.1016/j.avsg.2010.10.010
- 487 14. Lundgren F, Dahllof AG, Lundholm K, Schersten T, Volkmann R. Intermittent  
488 claudication. Surgical reconstruction or physical training? A prospective  
489 randomized trial of treatment efficiency. *Ann Surg*. 1989;209(3):346-355.  
490 doi:10.1097/00000658-198903000-00016
- 491 15. Murphy TP, Cutlip DE, Regensteiner JG, et al. Supervised exercise versus primary  
492 stenting for claudication resulting from aortoiliac peripheral artery disease: Six-  
493 month outcomes from the claudication: Exercise versus endoluminal  
494 revascularization (CLEVER) study. *Circulation*. 2012;125(1):130-139.  
495 doi:10.1161/CIRCULATIONAHA.111.075770
- 496 16. Whyman MR, Fowkes FGR, Kerracher EMG, et al. Is intermittent claudication  
497 improved by percutaneous transluminal angioplasty? A randomized controlled  
498 trial. *J Vasc Surg*. 1997;26(4):551-557. doi:10.1016/S0741-5214(97)70052-1
- 499 17. Perkins JMT, Collin J, Creasy TS, Fletcher EWL, Morris PJ. Exercise training versus  
500 angioplasty for stable claudication. Long and medium term results of a  
501 prospective, randomised trial. *Eur J Vasc Endovasc Surg*. 1996;11(4):409-413.

- 502           doi:10.1016/S1078-5884(96)80171-7
- 503   18.   RAND. 36-Item Short Form Survey (SF-36) Scoring Instructions. *Med. Outcomes*  
504           *Study 2–6* (2016).
- 505   19.   Chekroud, A. M. *et al.* The promise of machine learning in predicting treatment  
506           outcomes in psychiatry The promise of machine learning in predicting treatment  
507           outcomes in psychiatry. (2021) doi:10.1002/wps.20882.
- 508   20.   Al-Ramini, A. *et al.* Machine Learning-Based Peripheral Artery Disease  
509           Identification Using Laboratory-Based Gait Data. *Sensors* **22**, 7432 (2022).
- 510   21. Takallou MA, Fallahtafti F, Hassan M, Al-Ramini A, Qolomany B, Pipinos I, Myers S,  
511           Alsalem F. Diagnosis of disease affecting gait with a body acceleration-based  
512           model using reflected marker data for training and a wearable accelerometer for  
513           implementation. *Sci Rep.* 2024 Jan 11;14(1):1075. doi: 10.1038/s41598-023-  
514           50727-8. PMID: 38212467; PMCID: PMC10784467.
- 515   22.   Chicco, D., Tötsch, N. & Jurman, G. The Matthews correlation coefficient (MCC)  
516           is more reliable than balanced accuracy, bookmaker informedness, and  
517           markedness in two-class confusion matrix evaluation. *BioData Min.* 14, 13 (2021).
- 518   23Goulet-Pelletier, J.-C. & Cousineau, D. A review of effect sizes and their confidence  
519           intervals, Part I: The Cohen’s d family. *Quant. Methods Psychol.* **14**, 242–265  
520           (2018).
- 521   24.   Chen, S.-J. *et al.* Bilateral claudication results in alterations in the gait  
522           biomechanics at the hip and ankle joints. *J. Biomech.* **41**, 2506–2514 (2008).
- 523   25.   McDermott, M. M. *et al.* Measurement of walking endurance and walking velocity

- 524 with questionnaire: Validation of the walking impairment questionnaire in men and  
525 women with peripheral arterial disease. *J. Vasc. Surg.* **28**, 1072–1081 (1998).
- 526 26. Myers, S. A. *et al.* Claudication distances and the Walking Impairment  
527 Questionnaire best describe the ambulatory limitations in patients with  
528 symptomatic peripheral arterial disease. *J. Vasc. Surg.* **47**, 550–555 (2008).
- 529 27. Regensteiner, J. G., Steiner, J. F., Panzer, R. J. & Hiatt, W. R. Evaluation of  
530 walking impairment by questionnaire in patients with peripheral arterial disease. *J.*  
531 *Vasc. Med. Biol.* **2**, 142–152 (1990).
- 532 28. Brazier, J. E. *et al.* Validating the SF-36 health survey questionnaire: new  
533 outcome measure for primary care. *BMJ* **305**, 160–164 (1992).
- 534 29. Celis R, Pipinos II, Scott-Pandorf MM, Myers SA, Stergiou N, Johanning JM.  
535 Peripheral arterial disease affects kinematics during walking. *J Vasc Surg.* 2009  
536 Jan;49(1):127-32. doi: 10.1016/j.jvs.2008.08.013. Epub 2008 Nov 22. PubMed  
537 PMID: 19028062.
- 538 30. Atkinson, A. C., Riani, M. & Corbellini, A. The Box–Cox Transformation: Review  
539 and Extensions. *Stat. Sci.* **36**, (2021).
- 540 31. Ruxton, G. D. The unequal variance t-test is an underused alternative to Student’s  
541 t-test and the Mann–Whitney U test. *Behav. Ecol.* **17**, 688–690 (2006).
- 542 32. Hazra, A. & Gogtay, N. Biostatistics series module 3: Comparing groups:  
543 Numerical variables. *Indian J. Dermatol.* **61**, 251 (2016).
- 544 33. Menard, S. *Applied logistic regression analysis.* (Sage, 2002).
- 545 34. McDermott, M. M. *et al.* Prevalence and significance of unrecognized lower

- 546           extremity peripheral arterial disease in general medicine practice. *J. Gen. Intern.*  
547           *Med.* **16**, 384–390 (2001).
- 548   35.   Podgorelec, V., Kokol, P., Stiglic, B. & Rozman, I. Decision trees: an overview  
549           and their use in medicine. *J. Med. Syst.* **26**, 445–63 (2002).
- 550   36.   Battineni, G., Chintalapudi, N. & Amenta, F. Machine learning in medicine:  
551           Performance calculation of dementia prediction by support vector machines  
552           (SVM). *Informatics Med. Unlocked* **16**, 100200 (2019).
- 553   37.   Horst, F., Lapuschkin, S., Samek, W., Müller, K.-R. & Schöllhorn, W. I. Explaining  
554           the unique nature of individual gait patterns with deep learning. *Sci. Rep.* **9**, 2391  
555           (2019).
- 556   38.   Qutrio Baloch, Z., Raza, S. A., Pathak, R., Marone, L. & Ali, A. Machine Learning  
557           Confirms Nonlinear Relationship between Severity of Peripheral Arterial Disease,  
558           Functional Limitation and Symptom Severity. *Diagnostics* **10**, 515 (2020).
- 559