MASTER'S THESIS

OPTIMAL CARE FOR GERIATRIC HIP FRACTURE PATIENTS

An Interdisciplinary Perspective on Preoperative Decision-Making and Postoperative Rehabilitation

Michael Bui
March, 2023

zgt
Topzorg voor uw levenskwaliteit

UNIVERSITY OF TWENTE. | TECHMED CENTRE

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Submitted in partial fulfilment of the requirements for the degrees of Master of Science in Electrical Engineering and Master of Science in Interaction Technology

University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Master's programme: Electrical Engineering
Specialisation: Neurotechnology and Biomechatronics
Research group: Biomedical Signals and Systems

Master's programme: Interaction Technology
Research group: Human Media Interaction

Ziekenhuis Groep Twente (ZGT)
Department of Trauma Surgery

Graduation Committee

Dr. C.G.M. Groothuis-Oudshoorn	Health Technology and Services Research
Dr. A. Witteveen	Biomedical Signals and Systems
Dr. J.H. Hegeman	Biomedical Signals and Systems, ZGT
Dr. ing. G. Englebienne	Human Media Interaction
Dr. Y. Wang	Biomedical Signals and Systems
D. van Dartel, MSc	Biomedical Signals and Systems, ZGT
Dr. ir. B.J.F. van Beijnum	Biomedical Signals and Systems

Table of Contents

CHAPTER 1

General Introduction

1.1 Clinical Problem Statement

Hip fractures are a global health problem [1], mostly affecting older adults around the age of 80 years [2]. Due to the increasing life expectancy of the world population, an increasing incidence of hip fractures is anticipated in the upcoming years [3, 4]. Based on a worldwide estimate of 1.26-1.66 million hip fractures in 1990, epidemiological projections suggest that 2.6 million individuals will be affected annually by 2025. This number is expected to increase further to 4.5-6.26 million by 2050 [5, 6]. Hip fractures are acknowledged to be one of the most severe health problems affecting older adults [7, 8], being one of the most common causes of admission to acute orthopaedic wards [9].

In general, hip fractures have a poor survival prognosis as approximately one-third of the patients dies within one year following surgery [4]. It is postulated that the cause of death is attributable to degenerated physiological reserve [10], which is defined as "the potential capacity of a cell, tissue, or organ system to function beyond its basal level in response to alterations in physiological demands" [11, p. 492]. The poor physiological reserve is reflected by the high prevalence of multimorbidity amongst older adults [12], leading to an increased risk of developing postoperative complications such as pneumonia, pulmonary embolism, deep venous thrombosis, heart failure, myocardical infarction, and acute renal failure [13–18]. Therefore, the relatively poor health at baseline amongst older patients poses challenges for effective management of hip fractures.

Hip fracture management commences with fracture diagnosis through radiography, based on which adequate treatment procedures are considered. The fractures are classified as either extracapsular or intracapsular, where the latter is commonly subclassified based on the presence of displacement of the femoral neck [9]. In most cases, surgical treatment is recommended for each of these fractures [19]. Extracapsular and undisplaced femoral neck fractures are mostly managed with internal fixation, while displaced femoral neck fractures are managed with (hemi)arthroplasty to prevent avascular necrosis [19, 20].

In general, early surgery is advocated [19] since operative management effectively reliefs pain, allows for early mobilisation, and thereby prevents complications of immobilisation such as pressure ulcers [21]. Furthermore, a meta-analysis by Moja et al. [22] demonstrated that surgery within 24-48 hours significantly reduced mortality risk with an odds ratio of 0.74 (95% confidence interval: 0.67-0.81). Hence, surgery is considered to be the best treatment choice in worldwide practice for most hip fracture patients, yielding the highest likelihood of functional recovery and lowest mortality and complication rates [23].

However, for frail patients with a limited life expectancy, surgeons have begun to question the superiority of surgery over conservative treatment [24, 25]. Although current clinical guidelines favour surgical treatment based on prospects for functional recovery [26], these recovery-oriented objectives might not align with frail patients' personal preferences. Affirmatively, according to a systematic review examining patients' end-of-life care preferences, frail patients were more likely to decline invasive treatments than their age-matched controls [27]. Moreover, a recent study by Loggers et al. [28] found that conservative treatment was not inferior to surgery in terms of health-related quality of life (HRQoL) amongst frail institutionalised patients with limited life expectancy. Hence, surgery should not be a foregone conclusion for this patient population. However, the evidence-base underpinning that conservative treatment could be a satisfactory palliative care option is lacking [21]. Consequently, there is a paucity of concrete decision support for electing nonoperative management in current clinical guidelines.

Amongst patients for whom functional recovery is a viable surgical treatment objective, a substantial decline in health-related quality of life (HRQoL) is generally observed postoperatively [1, 29]. Gjertsen et al. [30] examined the differences between preoperative and postoperative HRQoL using the EQ-5D-3L instrument [31]. Amongst patients who did not report any HRQoL-related problems preoperatively, many experienced HRQoL-degenerating issues persisting over a one-year postoperative period, which concerned mobility (69.0%), self-care (40.7%), execution of usual activities (66.9%), pain or discomfort (65.7%), and anxiety or depression (36.5%). On the long-term, 29% of the older hip fracture patients experience lifelong functional disabilities [32]. In 10% of the cases, severe functional impairments even prohibit return to pre-fracture residence following rehabilitation [9]. Admission to a nursing home is perceived to be a major threat to the HRQoL, with many patients preferring earlier death over loss of their independence [33, 34]. Therefore, there is a pressing clinical need to further improve functional recovery during rehabilitation to enhance patients' HRQoL.

In current rehabilitation practices, a patient's functional recovery is assessed through clinimetric tests which examine patients' physical function, mobility, and cognition. Even though these insights play an important role in patient monitoring, clinimetric tests are conducted infrequently [35]. Consequently, important prognostic information about patients' recovery might be missed. As a result, necessary treatment adjustments may occur too late, potentially leading to a suboptimal recovery [36, 37]. Thus, hip fracture rehabilitation practices could strongly benefit from continuous monitoring strategies.

Multiple researchers have examined the usefulness of commercially available activity trackers for continuous ambulatory monitoring of geriatric hip fracture patients during

rehabilitation [35, 38]. They found that improvements in functional recovery measured through clinimetric tests, were positively correlated with the number of minutes that patients were physically active. Affirmatively, various studies have shown that physical activity during rehabilitation increases the likelihood of patients regaining their mobility and independence in activities of daily living (ADL) [39–42]. Hence, continuous ambulatory monitoring provides a promising means to gain more insights into a patient's restitution of physical activity [38], and thereby enables the detection of health deteriorations in a more timely manner. However, challenges persist as commercially available activity trackers do not reliably distinguish between slow physical activities and sedentary behaviours amongst rehabilitating hip fracture patients [35].

1.2 Thesis Contribution and Outline

It is evident that the hip fracture patient journey is complex. Along the way, it is pertinent that clinical decisions are supported by the best-available evidence, and that interventions are introduced in a timely manner to ensure satisfactory health outcomes. The optimality of clinical decisions and interventions, strongly depends on the specific needs of different subgroups in the hip fracture patient population [8, 43, 44]. On the one hand, frail patients with a limited life expectancy should be well-informed on the risks and benefits of palliative treatment alternatives to support decision-making. On the other hand, patients with sufficient physiological reserve should be supported to attain optimal functional recovery. Despite the focus areas being different, the overall objective of safeguarding patients' HRQoL remains the same. The contributions of this thesis are twofold.

The first part of this thesis focuses on decision support for frail geriatric hip fracture patients with a limited life expectancy. Chapter 2 provides a systematic review and meta-analysis of preoperative predictors for early mortality following hip fracture surgery. The results of this meta-analysis can help clinicians identify patients who are unfit for surgery. These patients in particular could benefit from conservative treatment. Chapter 3 reports on a clinical vignette study examining surgeons' treatment preferences for frail older adults with a hip fracture. In particular, it examines how individual patient attributes influence surgeons' risk perceptions and preferences for conservative treatment. Using a bottom-up expert-informed approach, the objective of the vignette study is to synthesise recommendations for the national guidelines on electing conservative treatment as a palliative care option.

The second part of this thesis focuses on continuous ambulatory monitoring systems to support hip fracture rehabilitation. Chapter 4 provides a literature review on common and best practices in the development of human activity recognition algorithms. The results of this literature review can be used by practitioners to make well-informed algorithmic design choices. Chapter 5 reports on the development of a human activity recognition algorithm for older adults, which can be applied to raw data obtained from wearable activity trackers. The results of this study yield a proof-of-concept of a continuous ambulatory monitoring system which could be utilised in geriatric hip fracture rehabilitation.

Finally, the main findings of this thesis are summarised in Chapter 6, accompanied by future perspectives.

CHAPTER 2

Systematic Review and Meta-Analysis of Preoperative Predictors for Early Mortality After Hip Fracture Surgery[1]

Abstract

Background: Hip fractures are a global health problem with a high postoperative mortality rate. Preoperative predictors for early mortality could be used to optimise and personalise healthcare strategies.
Objective: This study aimed to identify predictors for early mortality following hip fracture surgery.
Method: Cohort studies examining independent preoperative predictors for mortality following hip fracture surgery were identified through a systematic search on Scopus and PubMed. Predictors for 30-day mortality were the primary outcome, and predictors for mortality within one year were secondary outcomes. Primary outcomes were analysed with random-effects meta-analyses. Confidence in the cumulative evidence was assessed using the GRADE criteria. Secondary outcomes were synthesised narratively.
Results: 32 cohort studies involving 461,705 patients were included. Five high-quality evidence predictors for 30-day mortality were identified: age per year (OR: 1.06, 95% CI: 1.04-1.07), ASA score $\geq$ 3 (OR: 2.69, 95% CI: 2.12-3.42), male gender (OR: 2.00, 95% CI: 1.85-2.18), institutional residence (OR: 1.81, 95% CI: 1.31-2.49), and metastatic cancer (OR: 2.83, 95% CI: 2.58-3.10). Additionally, six moderate-quality evidence predictors were identified: chronic renal failure, dementia, diabetes, low haemoglobin, heart failures, and a history of any malignancy. Weak evidence was found for non-metastatic cancer.
Conclusion: This review found relevant preoperative predictors which could be used to identify patients who are at high risk of 30-day mortality following hip fracture surgery. For some predictors, the prognostic value could be increased by further subcategorising the conditions by severity.

Keywords: *older adults, hip fracture, mortality, risk factors, systematic review, meta-analysis*

[1] *Submitted* (Bui M, Nijmeijer WS, Hegeman JH, Witteveen A, Groothuis-Oudshoorn CGM, 2022)

2.1 Introduction

Hip fractures are a global health problem [1] with an increasing incidence due to the ageing population [3, 4]. According to epidemiological projections, 6.26 million individuals will be affected by hip fractures per year by 2050 [6]. Hip fractures are associated with an increased risk of mortality amongst older adults, with a cumulative 30-day mortality between 5-10% [20]. Over a 1-year postoperative period, it could accumulate up until approximately 30% [4].

Preoperative predictors for mortality following hip fracture surgery have been studied extensively [45–47]. Predictors for early mortality are particularly important, as they lie at the core of preoperative decision-making in clinical guidelines [26]. Preoperative prognostics could be used to better inform patients and family on the consequences of the different treatment alternatives, leading to better shared decision-making. This is particularly relevant for frail patients with a limited life expectancy who may experience a better quality of life if they do not undergo surgery [28]. Hence, shared decision-making could be leveraged to select a treatment that is optimal in terms of both clinical outcomes, and patients' personal values [48, 49]. This process ought to be supported by the best available evidence [50]. Meta-analyses can substantiate shared decision-making as they are one of the strongest resources in evidence-based medicine [51].

However, limitations in existing meta-analyses [45–47] impede effective support in shared decision-making. Firstly, evidence for early mortality predictors is scarce. Secondly, the relatively low number of included studies [52, 53] causes between-study heterogeneity underestimation [54], which makes significance testing more prone to false positives [55]. Although Bayesian meta-analyses could address this issue more adequately [52, 56–58], they have not been conducted in this field so far. Finally, to the best of our knowledge, none of the existing meta-analyses in this field have incorporated the Grades of Recommendation, Assessment, Development and Evaluation (GRADE) [59, 60] criteria to systematically assess the confidence in the cumulative evidence per predictor.

To support and improve evidence-based medicine for hip fracture patients, it is important to adequately reflect uncertainty in cumulative evidence. This will allow clinicians to assess the risk of early mortality more confidently, helping them to adequately inform their patients. The aim of this study is to conduct a meta-analysis, accompanied by GRADE assessments and (Bayesian) sensitivity analyses with respect to heterogeneity underestimation, to detect valid predictors for early postoperative mortality.

2.2 Method

This review was reported according to the PRISMA 2020 statement [61].

2.2.1 Search Strategy and Selection Criteria

The electronic databases Scopus and PubMed were searched from inception to 3 November 2021, using the search query as shown in Appendix A.1, Table A.1. Additionally, the Dutch Hip Fracture Audit (DHFA) was contacted for internal research reports.

2.2.2 Data Collection and Extraction

The title, abstract, and full-text screenings were performed by M.B., according to the exclusion criteria as described in Appendix A.1, Table A.2. The abstract and full-text screenings were independently verified by W.S.N. on a sample basis (70%). Disagreements were resolved through discussion. Study characteristics were extracted onto standardised tables containing author, year, country, study design, sample size, gender distribution, mean/median age, fracture types, treatment types, and mortality rates.

2.2.3 Outcome Measures

Adjusted odds ratios (ORs) and adjusted hazard ratios (HRs) of preoperative predictors for 30-day mortality following hip fracture surgery were primary outcomes. Independent predictors for mortality within one year were secondary outcomes.

2.2.4 Risk of Bias Assessment

The risk of bias of each included article was assessed by M.B. using the Quality In Prognosis Studies (QUIPS) tool [62]. A quarter of the articles were assessed independently by two reviewers (M.B, W.S.N.), who collectively refined the protocol to resolve ambiguities in the assessment criteria. Subsequently, the remaining articles were assessed by M.B. using the refined assessment criteria. The protocol can be found in Appendix A.2.

2.2.5 Data Synthesis

All predictors that were reported at least twice were synthesised in narrative summary tables [63], independent of whether they were reported as ORs or HRs. A minimum of three studies was set for quantitative synthesis and eligibility for pooling was based on consistency in variable definitions. ORs and HRs were meta-analysed separately for each of the predictors, using DerSimonian-Laird random-effects models [64] to accommodate for population and intervention heterogeneity [57, 65, 66]. Heterogeneity was quantified with the I^2 statistic and results were summarised with forest plots.

Sensitivity analyses were conducted with respect to publication bias, and between-study heterogeneity underestimation. The former was inspected with the trim-and-fill method [67] using the R_0^+ and L_0^+ algorithms [68], and the latter was inspected with the modified Knapp-Hartung method [69] and a Bayesian hierarchical model [70] (Appendix A.5). All analyses were performed with R version 4.1.2, using the *metafor* [71], *brms* [72], and *robvis* [73] packages.

2.2.6 Certainty of Evidence Assessment

Each pooled estimate was appraised using the GRADE criteria [59, 60] (Appendix A.4). When the quality of evidence was inconsistent across multiple pooled estimates of the same predictor, the quality of the pooled estimate based on most studies and patients was chosen for the final appraisal.

2.3 Results

2.3.1 Search and Included Studies

From the initial database yield of 1,869 articles, 139 were reviewed in full-text after assessing the eligibility based on titles and abstracts. Subsequently, an internal research report published by the DHFA was included and analysed. Reapplication of the exclusion criteria to the full texts yielded 100 articles for narrative synthesis, and 32 articles for meta-analysis. The selection process is shown in Figure 2.1.

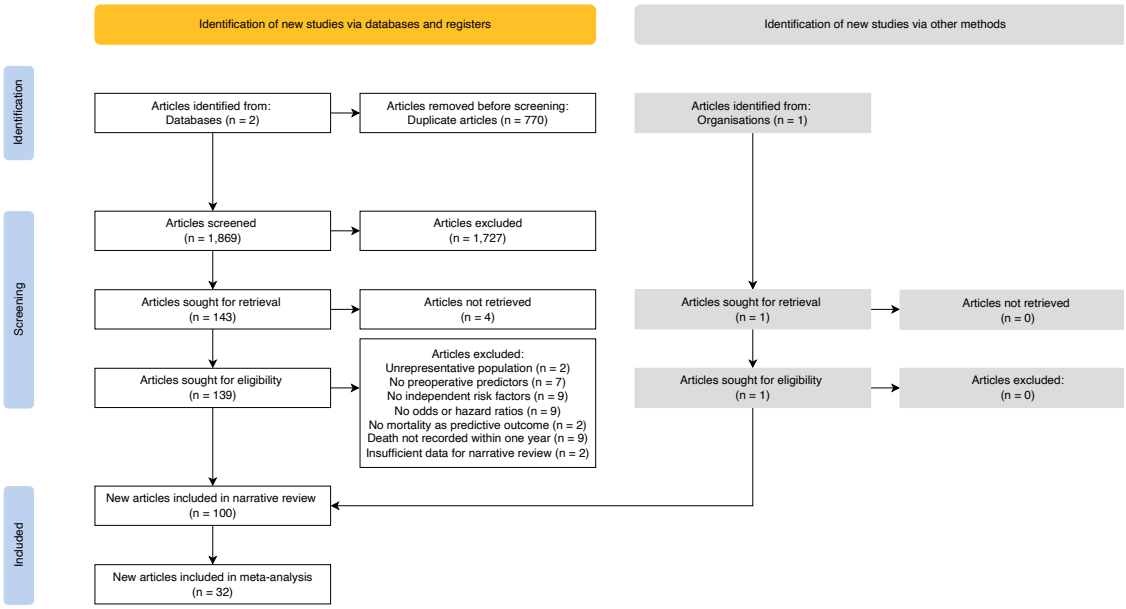

Figure 2.1: PRISMA flow diagram describing the identification, screening, and selection of articles.

A summary of the characteristics of the included studies is presented in Table A.3. Overall, early mortality was studied less frequently than late mortality. Predictors for inpatient, 30-day, and 1-year mortality were reported in 15, 35, and 62 studies respectively. Amongst the 32 studies reporting 30-day mortality predictors which were eligible for meta-analysis, involving 461,705 patients, one study did not report the 30-day mortality rate [74]. The median 30-day mortality rate and interquartile range across the remaining studies were 8.0% (6.5-9.6%).

2.3.2 Predictors for 30-day Mortality

An overview of all meta-analysed predictors for 30-day mortality is shown in Table 2.1, and forest plots of all high-quality evidence predictors are shown in Figure 2.2. The remaining forest plots are shown in Appendix A.6. None of the pooled evidence was downgraded for publication bias.

Table 2.1: Summary of findings table for the predictors of 30-day mortality following hip fracture surgery. The degree to which the studies included in the pooling procedures supported the association between the predictor and the increased risk of 30-day mortality is denoted by the direction of the association per study, where + denotes a significant result in favour of the association, 0 denotes a non-significant result in favour of the association, and - denotes a significant result refuting the association. Cases where any of the three directions are not applicable are denoted by N/A.

Predictor (measure)	**N**		**Association**	**Direction of association per study**			**Effect (95% CI)**	**GRADE**
	Patients	Studies		+	0	-		
Age per year (OR)	154,353	10	Greater 30-day mortality risk with advanced age.	[75–83]	[84]	N/A	1.06 (1.04-1.07)	High
ASA $\geq$ 3 (OR)	12,994	6	Greater 30-day mortality risk with increased ASA score.	[75, 80, 81, 85, 86]	[87]	N/A	2.69 (2.12-3.42)	High
ASA per point (OR)	5,394	3		[79, 82]	[88]	N/A	2.62 (2.21-3.12)	Moderate[a]
Chronic renal failure (OR)	248,872	3	Greater 30-day mortality risk with chronic renal failures.	[89, 90]	[76]	N/A	1.61 (1.11-2.34)	Moderate[b]
Dementia (OR)	389,185	6	Greater 30-day mortality risk of mortality with dementia.	[76, 89, 91–93]	[75, 94]	N/A	1.57 (1.30-1.90)	Moderate[c]
Dementia (HR)	29,929	3		[74, 95, 96]	N/A	N/A	1.47 (1.31-1.64)	High
Diabetes (OR)	378,573	4	Greater 30-day mortality risk with diabetes.	[89]	[76, 83, 91]	N/A	1.10 (1.01-1.21)	Moderate[b]
Gender (OR)	411,554	15	Greater 30-day mortality risk amongst males	[75, 76, 78, 79, 81, 82, 85, 88, 89, 91, 93, 97, 98]	[84, 99]	N/A	2.00 (1.85-2.18)	High
Gender (HR)	23,988	6		[95, 96, 100–102]	[103]	N/A	2.13 (1.94-2.34)	High
Hb per mmol/L (OR)	5,838	3	Greater 30-day mortality risk with lower Hb levels.	[75, 83]	[88]	N/A	1.37 (1.17-1.61)	Moderate[b]
Heart failure (OR)	384,312	5	Greater 30-day mortality risk with heart failures.	[76, 89–91, 104]	N/A	N/A	2.18 (1.25-3.82)	Moderate[c]
Institutional residence (OR)	6,638	3	Greater 30-day mortality risk with institutional residence.	[83, 105]	[75, 88, 92, 93]	N/A	1.81 (1.31-2.49)	High
Malignancy history	136,160	4	Greater 30-day mortality risk with a history of any malignancy.	[90, 91, 93]	[83]	N/A	2.39 (1.69-3.38)	Moderate[c]
Metastatic cancer (OR)	254,044	3	Greater 30-day mortality risk with metastatic cancer.	[76, 89, 104]	N/A	N/A	2.83 (2.58-3.10)	High
Non-metastatic cancer (OR)	249,192	3	Greater 30-day mortality risk with non-metastatic cancer.	[76, 89]	[92]	N/A	1.31 (1.11-1.56)	Low[bc]

GRADE Grading of Recommendations Assessment, Development and Evaluation, *CI* confidence interval, *OR* odds ratio, *HR* hazard ratio, *ASA* American Society of Anaesthesiologists physical status classification, *Hb* haemoglobin

[a] Downgraded by one level for risk of bias

[b] Downgraded by one level for imprecision

[c] Downgraded by one level for inconsistency

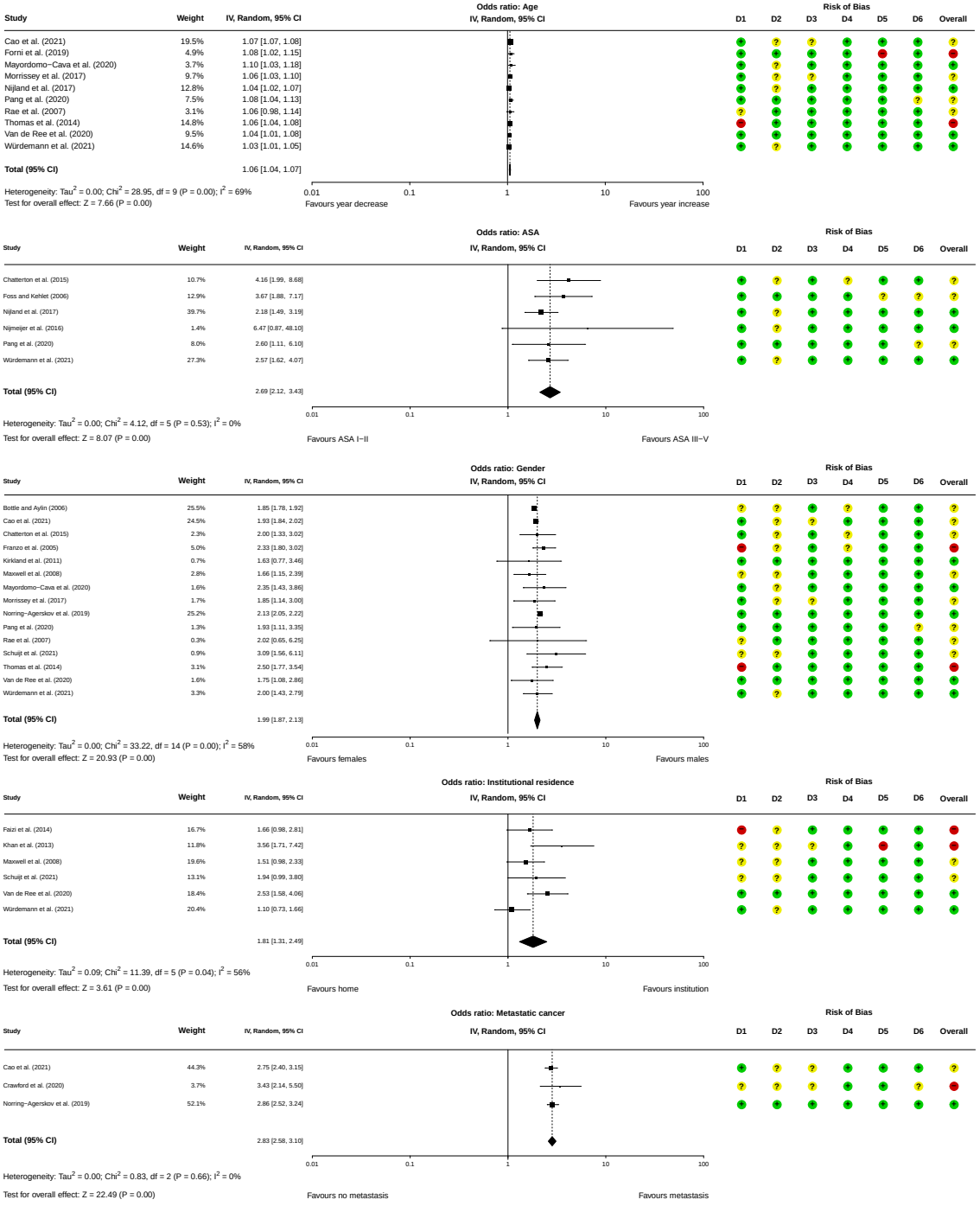

Figure 2.2: Forest plots of high-quality evidence predictors for 30-day mortality following hip fracture surgery. The right panel depicts the risk of bias assessments according to the bias domains of the Quality in Prognosis Studies tool i.e., study participation (D1), study attrition (D2), prognostic factor measurement (D3), outcome measurement (D4), study confounding (D5), and statistical analysis and reporting (D6). The risk of bias levels of low, moderate, and high, were colour-coded in green, yellow, and red respectively.

Age

Age was reported as both categorical and continuous variables. However, due to inconsistencies in the cut-off levels of age strata [85, 90, 93, 95, 96, 99–101], pooling was limited to studies reporting the influence of age per year increase. Analysis of 10 studies [75–84] including 154,353 patients provided high-quality evidence that a year increase in age increased the risk of 30-day mortality, with an OR of 1.06, 95% CI: 1.04-1.07. The forest plot in Figure 3 indicated that the pooled estimate overlapped with all 95% CIs, except for those reported by Cao et al. 1.07-1.08 and Würdemann et al. 1.01-1.05. Since the margin by which the CIs did not overlap was small, the interpretation of I^2 was deemed misleading. Therefore, it was decided against downgrading the quality of evidence for inconsistency, despite substantial heterogeneity ($I^2 = 69\%$).

American Society of Anaesthesiologists Score

American Society of Anaesthesiologists (ASA) scores were reported as both categorical and continuous variables across the studies. Amongst the reports of categorically treated ASA scores, two studies were excluded from pooling as there were insufficient data for the respective cut-off levels [76, 84]. Analysis of six studies [75, 80, 81, 85–87] including 12,994 patients provided high-quality evidence that individuals in ASA stratum III-V were at a greater risk of 30-day mortality than individuals in ASA stratum I-II, with an OR of 2.69, 95% CI: 2.12-3.42, $I^2 = 0\%$.

Similarly, analysis of three studies [79, 82, 88] including 5,394 patients provided moderate quality evidence that each unit increase in ASA score increased the risk of 30-day mortality with an OR of 2.62, 95% CI: 2.21-3.12, $I^2 = 0\%$. The quality of evidence was downgraded by one level for risk of bias as the cumulative weight of studies at high risk of bias was 71.6%.

Chronic Renal Failure

Renal failure was defined as end-stage renal failure (ESRF) [90], unspecified chronic renal failure (CRF) [76], moderate to severe CRF [89], and a joint stratum of acute renal failure (ARF) and early to end-stage CRF [91]. To keep the analysis homogeneous, instances of ARF were excluded from pooling.

Analysis of three studies [76, 89, 90] including 248,872 patients provided moderate-quality evidence that CRF increased the risk of 30-day mortality, with an OR of 1.61, 95% CI: 1.11-2.34, $I^2 = 50\%$. The quality of evidence was downgraded by one level for imprecision as both the Knapp-Hartung CI 0.52-5.23 and the Bayesian CrI 0.73-3.09 contained the null effect.

Dementia

Three studies did not report their dementia diagnoses [75, 76, 89], three studies reported on dementia in Alzheimer's disease [74, 91, 96], and one study reported on memory loss, (pre)senile and vascular dementias [95]. Two studies diagnosed dementia using an Abbreviated Mental Test Score ≤ 6 [92, 93], and one study diagnosed it with a

Hodkinson's abbreviated mental test score ≤ 6 [94]. Pooled estimates were not stratified by dementia diagnosis.

Analysis of three studies [74, 95, 96] including 29,929 patients provided high-quality evidence that dementia increased the risk of 30-day mortality, with a HR of 1.47, 95% CI: 1.31-1.64, $I^2 = 0\%$.

Similarly, analysis of seven studies [75, 76, 89, 91–94] including 389,185 patients provided moderate-quality evidence that dementia increased the risk of 30-day mortality, with an OR of 1.57, 95% CI: 1.30-1.90. The quality of evidence was downgraded for inconsistency due to substantial heterogeneity ($I^2 = 94\%$).

Diabetes

Analysis of four studies [76, 83, 89, 91] including 378,573 patients provided moderate-quality evidence that diabetes increased the risk of 30-day mortality, with an OR of 1.09, 95% CI: 1.01-1.18, $I^2 = 28\%$. The quality of evidence was downgraded for imprecision as both the Knapp-Hartung CI 0.96-1.25 and Bayesian CrI 0.84-1.43 contained the null effect.

Gender

Analysis of 15 studies [75, 76, 78, 79, 81–85, 88, 89, 91, 93, 97, 99] including 411,554 patients provided high-quality evidence that males were at a greater risk of 30-day mortality than females, with an OR of 1.99, 95% CI: 1.87-2.13, $I^2 = 58\%$.

Concordantly, analysis of six studies [95, 96, 100–103] including 23,988 patients provided high-quality evidence that males were at a greater risk of 30-day mortality than females, with a HR of 2.13, 95% CI: 1.94-2.34, $I^2 = 0\%$.

Haemoglobin

The influence of haemoglobin (Hb) was tested for anaemia (Hb $\leq$ 10 g/dL) [92, 93, 102], and per mmol/L decrease [75, 83, 88]. The former three studies comprised both ORs and HRs, causing an insufficiency in consistent data for pooling.

Analysis of three studies [75, 83, 88] including 5,838 patients provided moderate-quality evidence that a mmol/L decrease in Hb increased the risk of 30-day mortality, with an OR of 1.37, 95% CI: [1.17, 1.61], $I^2 = 40\%$. The quality of evidence was downgraded for imprecision as both the Knapp-Hartung CI 0.96-1.96 and Bayesian CrI 0.95-1.94 contained the null effect.

Heart Failure

Four studies did not report their HF diagnoses [76, 90, 100, 104], two studies diagnosed HFs using ICD-10 code I50 [89, 91], and one study included multiple hypertensive heart diseases in addition to ICD-10 code I50 [74]. Pooling was limited to studies reporting ORs since there were only two studies reporting HRs [74, 100].

Analysis of five studies [76, 89–91, 104] including 384,312 patients provided moderate-quality evidence that HF increased the risk of 30-day mortality, with an OR of 2.20, 95% CI: 1.28-3.78. The quality of evidence was downgraded for inconsistency due to substantial heterogeneity ($I^2 = 99\%$).

Malignancy

Four definitions of malignancies were found: history of any malignancy [83, 90, 91] excluding non-invasive skin cancer [93], non-metastatic cancer [76, 89, 92], and metastatic cancer [76, 89, 104]. Separate pooled estimates were computed for a history of any malignancy (excluding non-invasive skin cancer), non-metastatic cancer, and metastatic cancer.

Analysis of four studies [83, 90, 91, 93] including 136,160 patients provided moderate-quality evidence that a history of malignancy increased the risk of 30-day mortality, with an OR of 2.39, 95% CI: 1.69-3.38. The quality of evidence was downgraded by one level for inconsistency due to substantial heterogeneity ($I^2 = 61\%$).

Analysis of three studies [76, 89, 92] including 136,906 patients provided low-quality evidence that non-metastatic cancer increased the risk of 30-day mortality, with an OR of 1.17, 95% CI: 1.08-1.27. The quality of evidence was downgraded by one level for imprecision as both the Knapp-Hartung CI 0.99-1.73 and Bayesian CrI 0.95-1.86 contained the null effect, and by another level for inconsistency due to substantial heterogeneity ($I^2 = 80\%$).

Analysis of three studies [76, 89, 104] including 270,355 patients provided high-quality evidence that metastatic cancer increased the risk of 30-day mortality, with an OR of 2.83, 95% CI: 2.58-3.10, $I^2 = 0\%$.

2.3.3 Narrative Review Findings

The narrative review findings of predictors for postoperative mortality within one year, including 30-day mortality, are summarised in Table 2.2. Overall, the results were congruent with the meta-analysis. For institutional residence, however, the rate at which significant associations with mortality were found differed between short-term and long-term follow-ups. Table 2.1 showed that two-thirds of the studies contributing to the pooled estimate for institutional residence were insignificant. Upon including 4-month and 1-year follow-ups, two-thirds of the associations tested between institutional residence and mortality were significant.

Table 2.2: Summary of narrative review findings of adjusted odds and hazard ratios for the association between predictors and postoperative mortality within one year. + denotes a significant result in favour of the association, 0 denotes a non-significant result, and - denotes a significant result refuting the association. The final column depicts the relative frequency of significant associations per predictor.

Predictor	Association	Direction of association per study			Rel. freq. +
		+	0	-	
Age	Greater risk of mortality with advanced age.	[75–79, 81–83, 85, 86, 88–91, 93, 95–97, 99–102, 104, 106–144]	[84, 145–152]	N/A	61/70
Gender	Greater risk of mortality amongst males.	[14, 75, 76, 78, 79, 81–83, 85, 88, 89, 91, 93, 95–97, 100–102, 106, 108, 110, 112–115, 117, 118, 122–124, 128, 130, 131, 133–135, 137, 140, 141, 144, 147, 151, 153–156]	[84, 99, 103, 116, 120, 126, 127, 132, 138, 145, 149, 150, 152, 157–159]	N/A	48/64
ASA	Greater risk of mortality with increased ASA scores.	[75, 76, 79, 81, 82, 85–87, 104, 107, 108, 110, 112, 114–117, 131, 134, 136, 139, 141, 154, 160, 161]	[88, 109, 145, 147, 157, 158, 162]	N/A	25/32
Cognitive impairment	Greater risk of mortality with cognitive impairment.	[74, 76, 83, 89, 91–93, 95, 96, 114, 125, 127, 135, 139, 150, 151, 154, 157, 163]	[94, 109, 126, 149, 158, 159, 164, 165]	N/A	19/27
CCI	Greater risk of mortality with increased Charlson scores.	[74, 76, 94, 95, 97, 99, 102, 109, 117, 118, 122, 124, 140, 143, 147, 149, 158, 166, 167]	[126, 150, 153]	N/A	19/22
Malignancy	Greater risk of mortality with a history of malignancy.	[74, 76, 89–91, 93, 100, 107, 123, 125, 126, 128, 160, 164, 167, 168]	[83, 92]	N/A	18/20
Functional status in ADL	Greater risk of mortality with poorer functional status.	[78, 104, 107, 108, 113, 114, 126, 127, 134, 159]	[75, 94, 117, 121, 147, 158, 165]	N/A	10/17
Renal failure	Greater risk of mortality with of renal failures.	[14, 89–91, 100, 101, 122, 123, 133, 140, 169, 170]	[76, 155, 157, 168]	N/A	12/16
Congestive heart failure	Greater risk of mortality with congestive heart failures.	[74, 76, 89–91, 100, 113, 120, 123, 130, 133, 138, 164]	[76, 164]	N/A	13/15
Fracture type	Greater risk of mortality with extracapsular fractures vs. intracapsular fractures.	[76, 96, 122]	[79, 85, 103, 120, 124, 125, 132, 139, 145, 155, 158]	N/A	3/15
Institutional residence	Greater risk of mortality with pre-fracture institutional residence.	[83, 93, 100, 102, 105, 118, 123, 131, 153]	[75, 88, 92, 132, 158]	N/A	9/14
Haemoglobin	Greater risk of mortality with decreased haemoglobin levels (anaemia).	[75, 83, 93, 102, 127, 144, 171]	[88, 92, 147, 149, 157, 164]	N/A	7/13
Diabetes	Greater risk of mortality with diabetes.	[89, 91, 101, 128, 167]	[76, 83, 106, 120, 136, 155, 164]	N/A	5/12
BMI	Greater risk of mortality with lower BMI.	[88, 99, 104, 116, 127, 142, 160]	[147, 152]	N/A	7/9
Albumin	Greater risk of mortality with decreased albumin levels.	[94, 128, 132, 149, 159, 160, 169, 172]	N/A	N/A	8/8
Ischaemic heart disease	Greater risk of mortality with ischaemic heart disease.	[89, 91, 100, 119, 123, 130, 164]	[136]	N/A	7/8
COPD	Greater risk of mortality with COPD.	[74, 76, 120, 123, 133, 167, 169]	N/A	N/A	7/7
Number of comorbidities	Greater risk of mortality with an increased number of comorbidities.	[92, 93, 96, 137, 139, 162]	[115]	N/A	6/7
Mobility	Greater risk of mortality with poorer mobility.	[75, 86, 87, 131, 149, 158]	[152]	N/A	6/7
Myocardial infarction	Greater risk of mortality with myocardial infarction.	[90, 102, 123, 169]	[74, 76, 164]	N/A	4/7
Malnutrition	Greater risk of mortality with malnutrition.	[90, 117, 123, 131, 153, 165]	N/A	N/A	6/6
Cardiac arrhythmia	Greater risk of mortality with cardiac arrhythmia.	[78, 106, 123, 130, 133]	N/A	N/A	5/5
Electrolyte disorder	Greater risk of mortality with electrolyte disorder.	[78, 94, 123, 133, 173]	N/A	N/A	5/5
Bone mineral density	Greater risk of mortality with lower bone mineral density.	[112, 145, 152]	[75, 147]	N/A	3/5
Creatinine	Greater risk of mortality with higher creatinine levels.	[103, 120, 145, 147]	N/A	N/A	4/4
Hypertension	Greater risk of mortality with hypertension.	[168]	[106, 136]	[91]	1/4
Nottingham hip fracture score	Greater risk with higher Nottingham hip fracture scores.	[80, 87, 163]	N/A	N/A	3/4
Chronic liver disease	Greater risk of mortality with chronic liver disease.	[76, 102, 133]	N/A	N/A	3/3
Pneumonia	Greater risk of mortality with pneumonia.	[90, 123, 130]	N/A	N/A	3/3
Peripheral vascular disease	Greater risk of mortality with peripheral vascular disease.	[76, 133]	[74]	N/A	2/3
White blood cell count	Greater risk of mortality with lower white blood cell count.	[128]	[132, 174]	N/A	1/3
Hand grip strength	Greater risk of mortality with lower hand grip strength.	[127, 147]	N/A	N/A	2/2
Warfarin therapy	Greater risk of mortality with warfarin therapy.	[140, 175]	N/A	N/A	2/2

2.4 Discussion

This paper reports on the results of the first GRADE-compliant meta-analysis focusing on predictors of 30-day mortality following hip fracture surgery. In total, six high-quality evidence predictors were identified: age, gender, ASA classification, institutional residence, a history of malignancy, and metastatic cancer. Additionally, five moderate-quality evidence predictors were identified: CRF, dementia, diabetes, Hb, and HF. Finally, low-quality evidence was found for the influence of non-metastatic cancer.

To optimally use these findings in clinical practice, a few considerations must be made. Firstly, although a history of any malignancy is predictive of 30-day mortality, substantial heterogeneity exists in its prognostic value across studies ($I^2 = 61\%$). Better mortality risk predictions could be made if a distinction is made between non-metastatic and metastatic cancer, as the respective 95% CIs of 1.11-1.56 and 2.58-3.10 were distinct and showcased little variability. Although the necessity to make this distinction might seem straightforward, various 30-day mortality risk scores have not done this yet [83, 87, 176]. In accordance with the Charlson Comorbidity Index (CCI) [177], risk predictions should distinguish between non-metastatic and metastatic cancer to provide more accurate and personalised prognoses.

Secondly, CRF could manifest itself in different degrees of severity. Amongst the pooled studies, only one exclusively reported on the effect of ESRF [90]. Due to the low ESRF prevalence in 29/746 patients, the respective 95% CI was wide (1.05-10.01). Consequently, the meta-analysis did not reveal a need to stratify the risk estimate by severity of CRF as the individual 95% CIs overlapped by a sufficient margin to keep the between-study heterogeneity within acceptable bounds at $I^2 = 50\%$. However, larger studies with ESRF prevalences of 113/3,981 [123] and 886/44,419 patients [14] consistently reported larger risks of inpatient mortality with ORs of 6.70, 95% CI: 4.20-10.69 and 6.70, 95% CI: 3.57-12.58 respectively. Therefore, the pooled OR of 1.61 reported in this review is unlikely to be representative for patients with ESRF. Especially since CRF is highly prevalent amongst older adults [178], it becomes increasingly important to personalise prognoses based on the severity of CRF, rather than merely its presence or absence.

Thirdly, HFs might require a more careful operationalisation to be of better prognostic value. The pooled estimate reported in this review exhibited substantial unexplained heterogeneity ($I^2 = 99\%$). Even across studies which both resorted to ICD-10 code I50 for HF diagnosis [89, 91], the ORs differed substantially (95% CI: 1.54-1.73 vs 95% CI: 3.68-4.13). A disadvantage of ICD-10 code I50 is that it includes both HF with preserved ejection fraction and HF with reduced ejection fraction. Decreases in the left ventricular ejection fraction (LVEF) generally increase the risk of mortality [179]. It is postulated that the LVEF is an unobserved variable which could explain the high I^2 value. Therefore, future studies should acknowledge the varying degrees of severity in HFs and report the diagnoses in terms of the LVEF.

Several important limitations are noted. Some studies might have been overlooked since only two databases were searched for this review. Furthermore, the number of studies focusing on independent predictors of 30-day mortality is relatively limited, since most focus on more long-term prognoses. Consequently, the limited number of available studies restricted the use of additional methods to assess risk of publication bias more reliably, since funnel plots and Egger's test have very low power [180]. Hence, the conclusions drawn with respect to publication bias should be interpreted with caution.

Furthermore, the list of predictors is incomplete due to restrictions in pooling. Ischaemic heart disease was repeatedly associated with 30-day mortality, but could not be pooled as the results were a mix of ORs and HRs [89, 91, 100]. Additionally, inconsistency in reporting was identified as a systemic cause for incompleteness in the list of predictors. The CCI [76, 97, 99] and the number of comorbidities [92, 93, 137] were also repeatedly

found to be significant predictors of 30-day mortality. However, they could not be pooled since the cut-off levels by which patients were categorised were inconsistent.

Another issue induced by inconsistency in reporting manifested itself in the quality of pooled evidence. The pooled OR of Hb per mmol/L decrease was based on three studies instead of five due to inconsistent definitions for the influence of Hb. The respective quality of evidence was now downgraded for imprecision, which is postulated to have arisen due to a lack of power. Had all five studies been eligible for pooling, then sufficient power might have been attained to circumvent downgrading. Hence, future studies should establish which variable definitions and cut-off levels are most clinically relevant to the field of geriatric trauma surgery, e.g. by using the methods reported by Ogawa et al. [181], to improve consistency in reporting.

2.5 Conclusion

This study identified five high-quality, six moderate-quality, and one low-quality evidence predictors for 30-day mortality following hip fracture surgery based on preoperative data. Many of the published studies and widely used risk scores define predictors as the mere presence or absence of diseases. To provide better risk predictions, future studies should step away from such coarse definitions. According to the findings in this study, malignancies, CRFs, and HFs should be further subcategorised by severity to increase their prognostic value in prediction models. Hopefully, the results of this meta-analysis will enable clinicians to better identify patients who are at high risk of 30-day mortality. This information can be used to better inform patients on their prognosis, as one of the contributing factors which may lead to better shared decision-making in the preoperative phase.

CHAPTER 3

Surgeons' Treatment Preferences for Frail Geriatric Hip Fracture Patients: A Clinical Vignette Study

3.1 Introduction

In worldwide practice, operative treatment is considered to be superior over conservative treatment in terms of clinical outcomes for the majority of hip fracture patients [9, 23]. It is well-established that the mortality rate is significantly higher in conservatively treated patients than in operatively treated patients [98, 182, 183]. However, in case of frail older adults with a limited life expectancy, surgeons have started to question the superiority of surgery [24, 25]. Clinical guidelines often focus on functional recovery to pre-fracture levels [26], while patients with a limited life expectancy might prioritise their quality of life (QoL) instead [28]. In these cases, surgical overtreatment should be avoided due to its negative repercussions to patients and families, which include iatrogenesis and anxiety [184, 185]. There is increasing awareness that conservative treatments should be considered as a valid palliative care option more frequently amongst frail older adults [23, 25, 28, 186, 187].

Particularly amongst patients of advanced age with multiple physical and cognitive comorbidities, there is a pressing need for "counseling regarding prognosis for survival and recovery, and explicit discussions of goals of care" [188, p. 1279]. By properly informing frail patients on the available treatment options and examining how these align with their goals of care through shared decision-making (SDM) [49], patients and clinicians might come to the conclusion that conservative treatment is preferred. Affirmatively, a single-centre retrospective cohort study found that the percentage of patients electing nonoperative treatment increased significantly over the years (2.7% vs 9.1%) after implementing comprehensive geriatric assessments with SDM [189]. Still, uncertainties

regarding the optimal treatment choice might persist during SDM for complex patient cases [25]. Although surgeons increasingly acknowledge the added value of conservative treatment for frail patients, a paucity of decision support for palliative management in current clinical guidelines poses challenges for the preoperative decision-making process. Therefore, more decision support regarding the choice between operative and conservative treatment is required to optimise treatment plans for frail older adults.

Only a few studies have thus far investigated the motives behind electing conservative treatment. In most cases, conservative treatment was preferred when poor prognoses were anticipated for operative treatment, e.g. due to (chronic) comorbidities, poor functional status, and degenerating cognitive functioning [189, 190]. While these attributes could be used to identify patients who would not benefit from operative treatment, it remains a challenging task. Various prediction models for 30-day mortality following hip fracture surgery have been developed to identify patients who are unfit for operative treatment [83, 87, 176, 191, 192]. However, these models showcased moderate discriminative ability, making them premature for clinical practice. When data-driven approaches are not sufficiently reliable, domain experts should be consulted [193, 194]. Synthesis of clinicians' treatment preferences for various patient cases aids the understanding about which specific patients would benefit from which treatments [195].

The study presented here proposes a clinical vignette methodology to systematically elicit and analyse surgeons' treatment preferences for frail older hip fracture patients with limited life expectancy. This is a type of conjoint analysis (CA) [196, 197] in which the decision-making behaviours of medical experts are studied in various scenarios – so-called vignettes [198]. A vignette is defined as "a short, carefully constructed description of a person, object, or situation, representing a systematic combination of characteristics" [199, p. 128]. Given that clinicians' judgements of vignettes and their responses to real-life cases are sufficiently congruent [200], clinical vignette studies provide a means to reliably simulate and analyse complex decision-making processes in healthcare. The gained insights facilitate the understanding on which factors are influential in decision-making, to help inform clinical practices and policy development to support professional decision-making [201].

Stated differently, the clinical vignette methodology allows surgeons' treatment preferences to be studied in terms of the relative importance of individual patient attributes [196]. However, individual patient attributes also shape surgeons' overall perception of patients' early mortality risks. Capturing early mortality risk assessments is pertinent, since they could influence surgeons' perceptions of the benefit of operative treatment [24–26]. Therefore, the current study proposes to additionally elicit and synthesise surgeons' subjective probabilities of 30-day mortality following hip fracture surgery. This will be done using a structured expert judgement (SEJ) protocol [202].

To support preoperative decision-making for frail hip fracture patients with limited life expectancy, it is imperative to understand how patient characteristics and mortality risk perceptions affect surgeons' treatment preferences. Hence, this study aims to conduct a clinical vignette study and SEJ to systematically capture the expertise of trauma surgeons to synthesise recommendations for the national guidelines. To the best of our knowledge, this is the first ever clinical vignette study with an integrated SEJ.

3.2 Materials and Methods

3.2.1 Selection of Patient Attributes and Attribute Levels

Predictors for early mortality were chosen as primary attributes for the design of the vignettes, since conservative treatments are mostly reserved for patients with a limited life expectancy [26, 28]. Following the recommendations of [201], attributes were identified through a systematic review and meta-analysis (Chapter 2). To analyse surgeons' decision-making behaviours as comprehensively as possible, the vignettes were designed using the maximum number of attributes recommended in practice, i.e. 10 attributes [201].

All high-quality evidence predictors for 30-day mortality were extracted from the meta-analysis as attributes for the vignettes (age, gender, ASA score, institutional residence, and metastatic cancer). Amongst the five identified moderate-quality evidence predictors (see Table 2.1), only those for which confidence in the existence of a true significant association with mortality was expressed were selected (dementia, end-stage renal failure, and heart failure). To increase ecological validity, functional status in activities of daily living was included, as guidelines for preoperative decision-making are centred around functional recovery [28]. Finally, to enforce applicability to the study population of interest, fracture type was selected as an attribute.

A complete overview of the selected attributes and their dependencies is shown in Figure 3.1. The comorbidities and ASA scores showcased a direct dependency: ASA scores increase with the severity of diseases. Consequently, not all pairs of disease severity and ASA scores are logical to present in vignettes. Hence, attribute levels of comorbidities were defined such that they were maximally compatible with all ASA attribute levels chosen in this study. To keep the total number of vignettes low, the number of attribute levels was mostly restricted to two. Since it was anticipated that a dichotomy of health conditions and functional statuses could potentially be too coarse to inform decision-making, the vignettes were pilot tested with a surgical resident and medical specialist. Both clinicians agreed that it was not necessary to introduce additional attribute levels. An overview of the attribute levels along with the rationale behind the chosen definitions is depicted in Table 3.1.

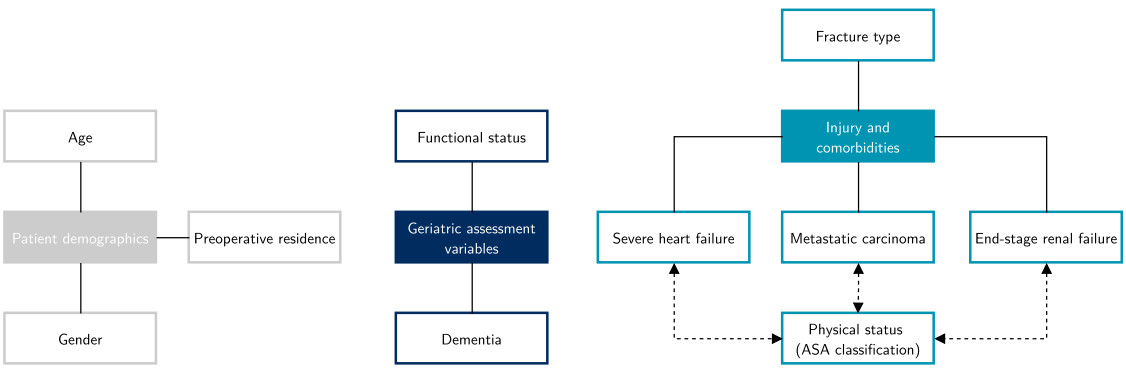

Figure 3.1: Overview of selected attributes, grouped into three overarching categories: patient demographics, geriatric assessment variables, and injury and comorbidities. Associations between attributes are depicted by dashed arrows.

Table 3.1: Overview of attributes and their levels, accompanied by the rationale for the level definitions.

Attribute	**Levels**	**Rationale**
Age	80-89 years old **OR** $\geq$ 90 years old	80 years was chosen as a lower bound, based on the average age of hip fracture patients. The cut-off between the two levels was based on the observation that complication risks and mortality rates differed significantly between octogenarians and nonagenarians [203].
Gender	Female **OR** male	-
Fracture type	Undisplaced femoral neck **OR** displaced femoral neck **OR** extracapsular	The invasiveness of the required surgical intervention differs between displaced and undisplaced femoral neck fractures. Most extracapsular fractures are treated with intramedullary nails in The Netherlands, omitting the need for more granular fracture type descriptions.
Physical status	ASA III **OR** ASA IV	It was anticipated that ASA I, II and V would not require decision support: all ASA I and II patients would be treated operatively [26], and all ASA V patients would be treated conservatively.
Severe heart failure	No severe heart failures (LVEF $\geq$ 30%) **OR** severe heart failure (LVEF $<$ 30%)	A moderate-to-severe reduction in LVEF is congruent with both ASA III and IV [204, 205]. The corresponding cut-off level of $\leq$ 30% was based on [179].
Metastatic carcinoma	No metastatic carcinoma **OR** metastatic carcinoma	The meta-analysis in Chapter 2 revealed a relatively weak association between non-metastatic cancer and 30-day mortality. Hence, no distinction was made between being free of cancer and having non-metastatic cancer.
End-stage renal failure	Not requiring dialysis **OR** requiring dialysis	Dialysis requirement complies with both ASA III and ASA IV classifications [204, 205]. Due to high renal failure prevalence amongst adults aged $\geq$ 80 years [178], no distinction was made between mild renal failures and absence of renal failures.
Preoperative residence	Home residence **OR** institutional residence	The meta-analysis in Chapter 2 did not reveal a need to further specify the type of care institutions. The reported prognostic values across different study settings were relatively homogeneous.
Functional status	No severe functional handicaps (Katz score 3-6) **OR** severe functional handicaps (Katz score 0-2)	Low pre-fracture functioning was a common cause for choosing conservative treatment [23, 189]. Hence, the extreme end of the Katz scale was chosen.
Dementia	No dementia **OR** dementia	Cognitive function declines with age [206], and the rate of decline even increases with age for vascular dementias [207]. A single level for dementia was thus thought to be sufficient to influence clinicians' decisions.

3.2.2 Experimental Design

The 10 attributes yielded a full factorial design comprising $2^9 3 = 1{,}536$ vignettes. However, one attribute level combination was deemed implausible: ASA III paired with metastatic cancer [208]. Hence, all vignettes containing this combination were removed from the full factorial design to reduce measurement errors [209], leaving a total of 1,152 vignettes. A D-optimal design [210] was generated from this subset with R version 4.0.2 using the *skpr* package [211]. The number of vignettes was minimised by inspecting the relative gain in D-efficiency upon increasing the number of vignettes over a range of 12 to 24. Based on these trials (Appendix B.1), a design comprising 16 vignettes was chosen (see Table 3.5) with a D-efficiency of 94.4% which implied near-orthogonality. The design's aliasing matrix showed a moderate correlation of 0.44 between ASA IV and metastatic cancer. The remaining correlations were weak (≤ 0.17) and mostly zero.

3.2.3 Questionnaire Design

The questionnaire comprised six sections. Firstly, surgeons were asked whether they were medical specialists or surgical residents, and how many years of working experience they had. Secondly, an explanation of all attribute levels was provided to prepare surgeons for the vignette study. Thirdly, surgeons were asked to recommend either operative or conservative treatments to each vignette, along with a statement on how certain they were about the optimality of their recommendation. Whenever surgeons recommended operative treatment, they were also asked whether they would perform surgery with curative or palliative intentions. Additionally, surgeons were asked to estimate the probability of 30-day mortality following hip fracture surgery for each vignette. Fourthly, an explanation on uncertainty specification was provided to prepare surgeons for quantile elicitation as part of the SEJ. Fifthly, surgeons were presented with the calibration questions of the SEJ, which are explained in more detail later. Finally, surgeons were asked what additional information in the vignette descriptions could have helped them to recommend treatments more confidently.

3.2.4 Data Collection

Surgical residents and medical specialists from the trauma surgery departments of three Dutch hospitals were surveyed between June and August 2022. The study was exempt from the Medical Research Involving Human Subjects Act, and it was approved by the ethical committee of Computer & Information Science of the University of Twente. All participants gave informed consent prior to participation.

3.2.5 Elicitation and Analysis of Treatment Preferences

The aim of the vignette study was to quantify the average impact of patient attributes on surgeons' treatment preferences. Most studies use a hierarchical logit with maximum likelihood (ML) estimation to accomplish this [212]. However, it was anticipated that the number of level-2 units (surgeons) would be too small for a hierarchical logit to unbiasedly estimate random- and fixed-effects [213, 214]. In case of few level-2 units, the assumption that ML estimates possess the property of asymptotic normality is violated

[215–217]. Consequently, the standard errors obtained from the Fisher information matrix are often underestimated, leading to an inflation of type I errors [214]. Although methods such as restricted maximum likelihood [218] and Kenward-Roger correction [219] have been proposed to address the aforementioned issues [214], these remedies are premature for hierarchical models with discrete outcomes [215]. Bayesian model adaptations, on the other hand, can overcome the small sample limitations of the hierarchical logit through a careful choice of (weakly) informative priors [213, 215, 217, 220, 221]. Hence, the data were analysed with a hierarchical Bayesian logit model with random-effects intercepts. Analyses were performed and reported according to the WAMBS-checklist [222]. For a surgeon s_i who judged N vignettes, the model is described by equation (3.1),

$$\begin{cases} \phi(\boldsymbol{p}_{s_i}) = \boldsymbol{X}\boldsymbol{\beta} + u_{s_i}\mathbf{1} + \boldsymbol{\epsilon}_{s_i} \\ \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ u_{s_i} \sim \mathcal{N}(0, \sigma_u^2) \\ \sigma_u^2 \sim \mathcal{IW}(\psi, \nu) \\ \boldsymbol{\epsilon}_{s_i} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_N) \end{cases} \tag{3.1}$$

where $\phi(\boldsymbol{p}_{s_i})$ is an $N \times 1$ latent utility vector describing the perceived benefit of conservative treatment, with ϕ being the logit link function and $\boldsymbol{p}_{s_i}$ being a probability vector, $\boldsymbol{X}$ is the $N \times K$ design matrix, $\boldsymbol{\beta}$ is the $K \times 1$ vector of part-worth utilities approximated by log odds ratios, $\mathbf{1}$ is an $N \times 1$ vector of ones, u_{s_i} is the random intercept of surgeon s_i, σ_u^2 is the variance of u_{s_i}, and $\boldsymbol{\epsilon}_{s_i}$ is the $N \times 1$ random error vector.

$\boldsymbol{\beta}$ was modelled with a Gaussian prior since log odds ratios follow an approximate normal distribution [223]. Since early mortality risk is the primary reason for electing conservative treatment [26], the effect sizes of the meta-analysis in Chapter 2 were used to estimate $\boldsymbol{\beta}$ (see Table 3.2). The subject-specific residual u_{s_i} was modelled using a Gaussian prior with a mean of zero and variance σ_u^2. The latter was specified with an inverse Wishart hyperprior, which is common choice due to its conjugacy with the Gaussian distribution [224–226]. Since no prior information was available on random-effects, the most conservative prior was chosen with a 1×1 scale matrix $\psi = 1$ and $\nu = 1$ degrees of freedom [225]. Finally, $\boldsymbol{\epsilon}_{s_i}$ was modelled using a Gaussian prior with a mean vector equal to the zero vector and a covariance matrix equal to the identity matrix $\boldsymbol{I}_N$ [227].

The posterior distributions for each part-worth utility were estimated via Markov Chain Monte Carlo (MCMC) sampling [228] using a blocked Gibbs sampler [229]. For each part-worth utility, 15,000 posterior samples were drawn after a burn-in phase of 1,000 samples. Point estimates were obtained by computing the posterior means [230]. The model was implemented in R version 4.0.2, using the *MCMCpack* [231] package. Finally, the relative importance of each attribute was computed using the coefficient range method [232]. Let V_q denote the importance of an attribute q, defined as the maximum range of part-worths across all attribute levels of q. Then, the relative importance of q is computed by normalising V_q over the sum of all Q importance values (3.2).

$$I_q = \frac{V_q}{\sum_{i=1}^{Q} V_i} \times 100\% \tag{3.2}$$

Table 3.2: Overview of prior specifications expressed on a logarithmic scale. Unless stated differently, all odds ratios (ORs) serving as secondary evidence for prior specifications were obtained from the meta-analysis described in Chapter 2.

Parameter	Distribution	Specification	Prior type	Background knowledge
β_0	Normal	$\mathcal{N}(-2.75, 1)$	Weakly informative	3% of the Dutch patients is treated conservatively [75]. As the vignettes exclude ASA I-II, β_0 was expected to be slightly higher. The prior yields a mean probability of 6.0% (95% CrI: 0.9-31.2%) in favour of conservative treatment for the null model.
β_{gender}	Normal	$\mathcal{N}(0.09, 1)$	Weakly informative	Male gender is a high-quality evidence predictor for 30-day mortality. However, it was deemed unlikely that this would be reflected in surgeons' treatment preferences. Hence, the informativeness of the prior was decreased, yielding a mean OR of 1.1 (95% CrI: 0.15-7.80) in favour of conservative treatment.
$\beta_{\text{extracapsular}}$	Normal	$\mathcal{N}(0.09, 1)$	Weakly informative	Compared to undisplaced femoral neck fractures, extracapsular fractures have a higher postoperative anaemia incidence [233]. Due to the lack of strong evidence for increased mortality risk [47], a small mean OR of 1.1 (95% CrI: 0.15-7.80) in favour of conservative treatment was assumed.
β_{DFN}	Normal	$\mathcal{N}(0.18, 1)$	Weakly informative	Displaced femoral neck fractures require more invasive surgical intervention than their undisplaced counterparts. As quantitative evidence was lacking, a small mean OR of 1.2 (95% CrI: 0.17-8.51) in favour of conservative treatment was assumed.
β_{ASA}	Normal	$\mathcal{N}(0.69, 1)$	Informative	ASA scores increase 30-day mortality risk with an OR of 2.62 (95% CI: 2.21-3.12) per point increase. During the vignette study pilot test, a surgical resident expressed indifference towards ASA scores, due to subjectivity of the scoring system [234]. A relatively wide prior was chosen to reflect uncertainty in the influence of ASA scores, with a mean OR of 2.0 (95% CrI: 0.5-7.99) in favour of conservative treatment.
β_{heart}	Normal	$\mathcal{N}(0.69, 0.5)$	Informative	Heart failure increases the risk of 30-day mortality with an OR of 2.18 (95% CI: 1.25-3.82). The prior yields a mean OR of 2.0 (95% CrI: 0.50-7.98) in favour of conservative treatment.
$\beta_{\text{metastasis}}$	Normal	$\mathcal{N}(0.92, 0.3)$	Informative	Metastasis increases 30-day mortality risk with an OR of 2.83 (95% CI: 2.58-3.10). Informativeness of the prior was increased due to high quality of the evidence and the narrow CI width. The prior yields a mean OR of 2.5 (95% CrI: 0.85-7.32) in favour of conservative treatment.
β_{ESRF}	Normal	$\mathcal{N}(0.79, 0.5)$	Informative	Chronic renal failure increases the risk of 30-day mortality with an OR of 1.61 (95% CI: 1.11-2.34). Cohort studies have shown that inpatient mortality risk is even higher for ESRF (95% CI: 3.57-12.58) [14]. The prior yields a mean OR of 2.2 (95% CrI: 0.55-8.81) in favour of conservative treatment.
$\beta_{\text{institution}}$	Normal	$\mathcal{N}(0.47, 0.5)$	Informative	Institutional residence increases the risk of 30-day mortality with an OR of 1.81 (95% CI: 1.31-2.49). The prior yields a mean OR of 1.6 (95% CrI: 0.40-6.42) in favour of conservative treatment.
$\beta_{\text{functional}}$	Normal	$\mathcal{N}(0.47, 0.7)$	Informative	It was assumed the effect size of severe functional handicaps was similar to that of institutional residence. However, due to the lack of quantitative evidence, a slightly wider prior was specified with a mean OR of 1.6 (95% CrI: 0.31-8.26).
β_{dementia}	Normal	$\mathcal{N}(0.34, 0.5)$	Informative	Dementia increases the risk of 30-day mortality with an OR of 1.57 (95% CI: 1.30-1.90). The prior yields a mean OR of 1.4 (95% CrI: 0.35-5.60).
u_{s_i}	Normal	$\mathcal{N}(0, \sigma_u^2)$	Uninformative	N/A
σ_u^2	Inverse Wishart	$\mathcal{IW}(1, 1)$	Uninformative	N/A
$\boldsymbol{\epsilon}_{s_i}$	Normal	$\mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$	Uninformative	N/A

3.2.6 Convergence Diagnostics and Sensitivity Analysis

Trace and autocorrelation plots were inspected for MCMC convergence. Signs of good convergence were trendless traces and rapid decays of the autocorrelation functions. To test whether the Markov chains had truly converged or only converged locally, the number of posterior samples was doubled to 30,000 samples. Subsequently, Geweke's convergence test [235] was conducted to assess whether stationarity between the first 15,000 samples and the last 15,000 samples could be assumed. P-values above 0.05 were indicative of healthy convergence. To determine whether the resulting posteriors were sufficiently smooth, histograms of the posterior draws were inspected.

Finally, to assess the extent to which subjectivity in the prior specifications affected the odds ratios (ORs), the hierarchical Bayesian logit model was re-evaluated with flattened Gaussian priors, i.e. $\mathcal{N}(0, 2)$, for each β (see Figure 3.2). The influence of priors was considered (1) small if the relative deviation Δ (3.3) was at most 10% and the substantive results remained the same, (2) moderate if $10\% < \Delta \leq 20\%$ and the substantive results remained the same, and (3) large otherwise.

$$\Delta = \frac{|\mathrm{OR}_{\text{informative}} - \mathrm{OR}_{\text{noninformative}}|}{\mathrm{OR}_{\text{informative}}} \times 100\% \tag{3.3}$$

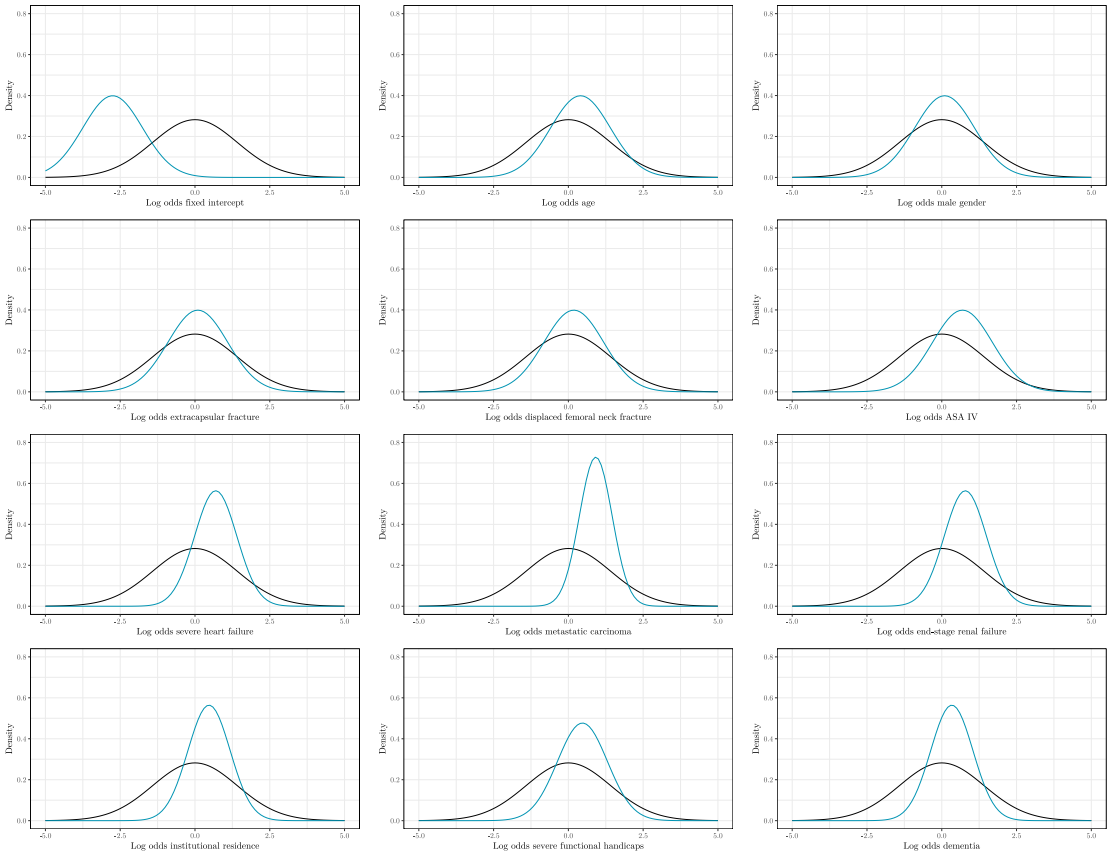

Figure 3.2: Overview of prior distributions. The probability density functions plotted in blue represent the priors which were proposed for the primary regression analysis. The probability density functions plotted in black depict more diffuse prior configurations which were used in sensitivity analyses.

3.2.7 A Priori Power Analysis and Sample Size Calculations

Health policy recommendations based on non-significant outcomes should not be made without considering whether the study had sufficient power to detect small yet meaningful effects [236]. Hence, an a priori power analysis was conducted using Monte Carlo simulations [237], as described in Algorithm 1.

Algorithm 1: Power Analysis Through Monte Carlo Simulations

Result: Matrix $\boldsymbol{\Pi}$, where element $\pi_{i,k}$ is the power estimate of the k-th logit model coefficient, for the i-th surgeon cohort size

Initialise empty vector of binary treatment choices $\boldsymbol{y}$
Initialise logit model coefficients $\beta_0, \beta_1, ..., \beta_K$ to the expected effect sizes
Initialise design matrix $\boldsymbol{X} \in \mathbb{R}^{M \times K}$ to the D-optimal design
Initialise vector of surgeon cohort sizes $\boldsymbol{s}$
Initialise number of simulations n
Initialise empty matrix of p-values $\boldsymbol{P}$

for *each* $s_i \in \boldsymbol{s}$ **do**
  repeat
    repeat
      for *each row in* $\boldsymbol{X}$ **do**
        $y_j \leftarrow \text{Bernoulli}\left(\frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K\right)}\right)$
        Append y_j to $\boldsymbol{y}$
      end
    until s_i *response sets have been generated for* $\boldsymbol{X}$
    Compute logit model $\hat{f}(\boldsymbol{X}, \boldsymbol{y})$
    $\boldsymbol{y} \leftarrow \{\emptyset\}$
    Compute p-value vector $\boldsymbol{p}$ corresponding to $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_K$ in $\hat{f}(\boldsymbol{X}, \boldsymbol{y})$
    Append $\boldsymbol{p}$ to the columns of $\boldsymbol{P}$
  until n *simulations have been performed*
  $\pi_{i,k} \leftarrow \frac{1}{n} \sum_{q=1}^{n} \mathbb{1}_{\{p_{q,k} < 0.05\}}, \forall k \in \{1, 2, ..., K\}$
end

Contrary to rules of thumb which are commonly used for CA in healthcare [238], simulations provide estimates which capture both the analytic outcome model and the study design [239]. For simplicity, random-effects were omitted from the simulations to obtain a rough estimate for the required sample size. Hence, a logit model was used as analytic outcome model in the power simulations. The prior beliefs of the effect sizes were kept consistent with the prior distribution means assumed in the Bayesian hierarchical model, as outlined in Table 3.2. The resulting power curves shown in Figure 3.3 indicated that approximately 55 respondents were required to attain a power above 60% for 8/11 attribute levels. Significant results for the remaining attribute levels (extracapsular fracture, displaced femoral neck fracture, and male gender) were expected to be unattainable due to low power.

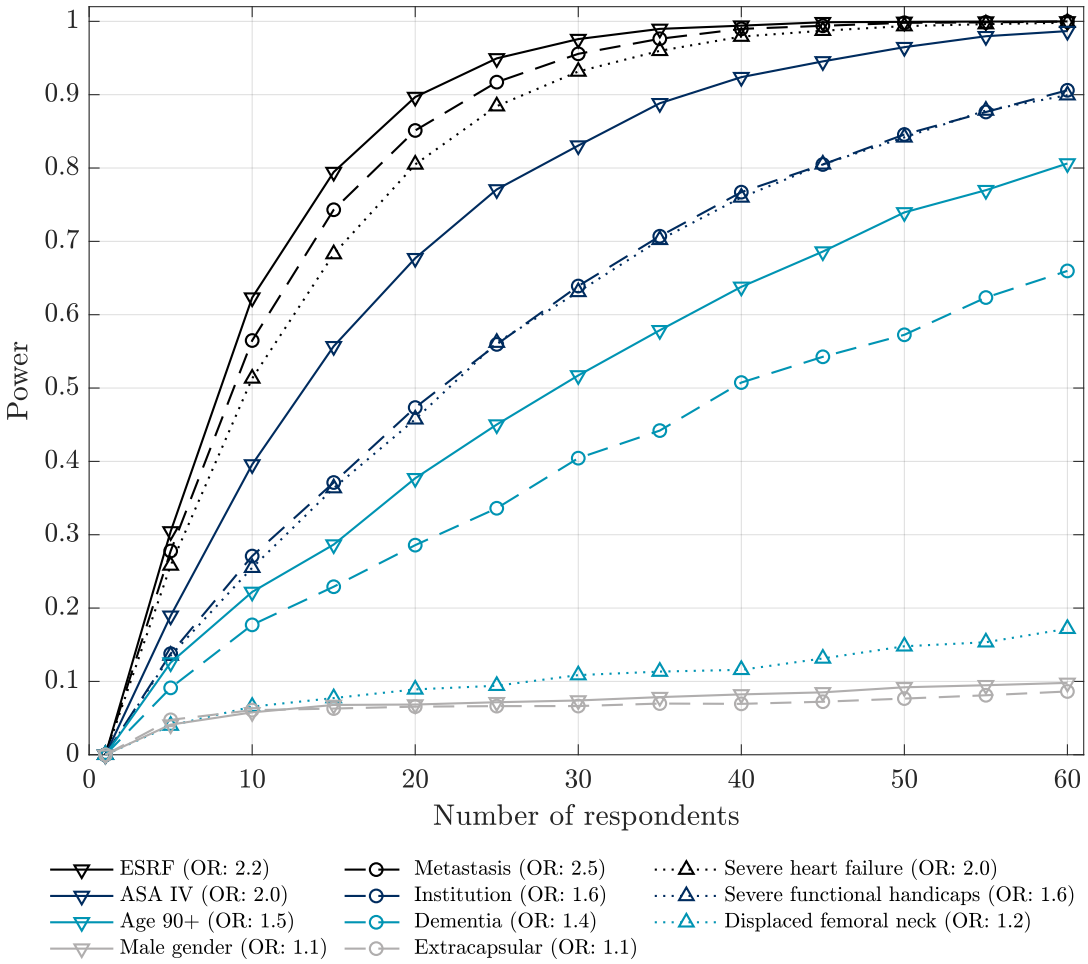

Figure 3.3: Power curves for the attribute levels used in the vignettes. The respective odds ratios (ORs) which were assumed during the power calculations, are listed behind each attribute level.

3.2.8 Elicitation and Analysis of Risk Perceptions

The goal of the SEJ was to elicit and aggregate surgeons' 30-day mortality risk perceptions of frail older adults undergoing hip fracture surgery. Risk perceptions were elicited by asking surgeons to estimate the probabilities that the patients described in the vignettes would die within 30 days following surgery. Expert elicitation was performed using Cooke's Classical Model (CM) [202], which is exemplar in the field of SEJs [240] as it is the only elicitation procedure with objective empirical control on expert scoring [241]. The CM enforces empirical control by first scoring how statistically accurate and informative surgeons are in the estimation of verifiable variables, prior to aggregating their judgements on unknown variables. Surgeons with higher scores are assigned higher performance-based weights in the aggregation, to obtain the best estimate of the unknown target variable.

Calibration questions (CQs) were used as instruments to measure surgeons' performances. CQs involve the estimation of so-called seed variables, which refer to quantities with a close relation to the target variable. In this case, 30-day mortality prevalence percentages amongst subpopulations of hip fracture patients were chosen as seed variables. Surgeons are not expected to know the exact percentages, but they should be able to capture the

seed variables' true realisations reliably based on their expertise by defining adequate credible intervals (CrIs). The 5%, 50%, and 95% quantiles, i.e. q_5, q_{50} and q_{95}, were chosen for CrI elicitation as this is most common practice in SEJs [242, 243].

Structured Expert Judgement Instruments

For each vignette, the following target question was posed: *"According to you, what is the probability that a patient with these characteristics would die within 30 days after hip fracture surgery?"*. Surgeons were asked to choose a probability bin from the set {[0,0.1), [0.1,0.2), ..., [0.9,1.0]} which reflected their beliefs best, which is the most common response format for discrete event probabilities in SEJs [194, 202]. The middle value of each bin functioned as a point estimate for pooling later in the analysis.

CQs were based on 30-day mortality data from the Dutch Hip Fracture Audit Taskforce Indicators (DHFA-TFI) group [75], which described a total of 7,506 patients. The CQs involved the estimation of 30-day mortality prevalence following hip fracture surgery, based on preoperative characteristics. To ensure similarity with the target questions, prevalences were extracted from patient subgroups which were age-matched with those described in the vignettes ($\geq$ 80 years). In addition to age matching, overlapping attributes between the vignettes and the DHFA-TFI data were incorporated in the patient descriptions for the CQs as well. These included gender, fracture type, dementia, functional status in ADL, ASA scores, and institutional residence. Since these characteristics were not sufficient to construct 14 diverse CQs, mobility, malnutrition, and anaemia were included as additional attributes. An example of a CQ is: *"How many percent of the hip fracture patients aged **90 years or older**, who were **mobile without walking aids** and **did not have dementia**, died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?"* An overview of all CQs is shown in Appendix B.2.

The true realisations of the seed variables could not be obtained directly since there were missing data. Information on 30-day mortality was missing for 19.5% of the 7,506 patients. Missing entries were imputed with Multiple Imputation by Chained Equations (MICE) [244]. Preoperative and perioperative risk factors for mortality were chosen as variables for imputation. Imputations of numerical variables and categorical variables with an ordinal scale were configured with predictive mean matching. Imputations of the remaining categorical variables were configured with binary and multinomial logistic regression. MICE was used to create 20 imputed data sets [244], from which the 30-day mortality percentages were extracted and pooled using Rubin's rules [245]. With the true realisations of the seed variables available, surgeons' performances could be measured using two scoring metrics: the calibration score and the information score.

Calibration Score

The calibration score evaluated how statistically accurate surgeons' CrIs captured the true realisations of seed variables. Calibration was measured by examining the interquantile ranges [0, 0.05], (0.05, 0.50], (0.50, 0.95], (0.95, 1.0] of surgeons' elicited quantiles across all CQs. Specifically, an empirical distribution $\boldsymbol{e}(s_i) = (e_1(s_i), e_2(s_i), e_3(s_i), e_4(s_i))$ of a surgeon s_i's interquantile ranges was constructed by computing the proportions of true realisations falling within each interquantile range. By definition of quantiles, a surgeon

s_i is said to be well-calibrated if the empirical distribution $\boldsymbol{e}(s_i)$ resembles the theoretical distribution $\boldsymbol{p}$ = (0.05, 0.45, 0.45, 0.05). The test statistic $2MI(\boldsymbol{e}(s_i), \boldsymbol{p})$ (3.4) was used to test the null hypothesis of $H_0 : \boldsymbol{e}(s_i) = \boldsymbol{p}$, where M is the number of CQs and $I(\boldsymbol{e}(s_i), \boldsymbol{p})$ is the Kullback-Leibler divergence. $I(\boldsymbol{e}(s_i), \boldsymbol{p})$ measures the discrepancy between $\boldsymbol{e}(s_i)$ and $\boldsymbol{p}$ and yields a value of zero if they are identical.

$$2MI(\boldsymbol{e}(s_i), \boldsymbol{p}) = 2M \sum_{l=1}^{4} e_l(s_i) ln \frac{e_l(s_i)}{p_l} \tag{3.4}$$

The calibration score is defined as the p-value of the hypothesis test examining whether surgeons are well-calibrated. The test statistic $2MI(\boldsymbol{e}(s_i), \boldsymbol{p}) \sim \chi^2$ with three degrees of freedom as $M \to \infty$. Under these distributional assumptions, the p-value corresponding to this test statistic is given by equation (3.5), where $F(.)$ is the cumulative distribution function (CDF) of the Chi-squared random variable. The closer $Cal(s_i)$ is to 1, the better a surgeon is calibrated and the more likely it is that $H_0 : \boldsymbol{e}(s_i) = \boldsymbol{p}$ is true.

$$Cal(s_i) = 1 - F(2MI(\boldsymbol{e}(s_i), \boldsymbol{p})) \tag{3.5}$$

Information Score

Information in surgeons' judgements referred to the degree to which their subjective probability distributions were concentrated. This partially reflects how certain surgeons are about their assessments, but it mostly indicates whether surgeons deem some values more likely to be true than others. Based on the elicited quantiles, CDFs were constructed per CQ for each surgeon. However, since surgeons only specified their distribution partially through 90% CrIs, the CDFs could not be constructed directly as their supports were unknown. Hence, the support was defined manually using the intrinsic range $[L^*, U^*]$ = [0, 100], which is suitable for seed variables expressed in percentages [246].

Subsequently, informativeness was quantified by comparing surgeons' CDFs (3.6) to a uniform background measure (3.7) across the four interquantile ranges. The motivation behind choosing a uniform background measure was that the presence of any information should be quantified relative to the least informative background [202].

$$F(x) = \begin{cases} 0.05, & \text{for } x \in [L^*, q_5] \\ 0.45, & \text{for } x \in (q_5, q_{50}] \\ 0.45, & \text{for } x \in (q_{50}, q_{95}] \\ 0.05, & \text{for } x \in (q_{95}, U^*] \end{cases} \tag{3.6}$$

$$B(x) = \begin{cases} \dfrac{q_5 - L^*}{U^* - L^*}, & \text{for } x \in [L^*, q_5] \\ \dfrac{q_{50} - q_5}{U^* - L^*}, & \text{for } x \in (q_5, q_{50}] \\ \dfrac{q_{95} - q_{50}}{U^* - L^*}, & \text{for } x \in (q_{50}, q_{95}] \\ \dfrac{U^* - q_{95}}{U^* - L^*}, & \text{for } x \in (q_{95}, U^*] \end{cases} \tag{3.7}$$

The information score was then determined by computing the Kullback-Leibler divergence between a surgeon's CDF and the uniform CDF applied to the intrinsic range. From equations (3.6) and (3.7) then followed that the information score of a surgeon s_i for CQ j was given by equation (3.8). The total information score of a surgeon s_i was obtained by averaging the information scores across all M CQs (3.9).

$$\begin{aligned} I_j(e_i) = 0.05\, ln\left(\frac{0.05}{q_5 - L^*}\right) + 0.45\, ln\left(\frac{0.45}{q_{50} - q_5}\right) + 0.45\, ln\left(\frac{0.45}{q_{95} - q_{50}}\right) \\ + 0.05\, ln\left(\frac{0.05}{U^* - q_{95}}\right) + ln(U^* - L^*) \end{aligned} \tag{3.8}$$

$$I(s_i) = \frac{1}{M}\sum_{j=1}^{M} I_j(s_i) \tag{3.9}$$

Performance-based Weighting

The calibration scores and information scores were combined into performance-based weights, normalised across all K surgeons, using equation (3.10).

$$w(s_i) = \frac{Cal(s_i)I(s_i)}{\sum_{k=1}^{K} Cal(s_k)I(s_k)} \tag{3.10}$$

For each vignette, the target variables, i.e. the probabilities of 30-day mortality, were aggregated using linear opinion pooling with performance-based weights. It should be noted that the target variables described the occurrence of a binary discrete random variable (DRV), while the calibration and information scores were configured for CQs involving continuous random variables (CRVs). Traditionally, target variables and seed variables should either both be CRVs, or both concern DRVs [246]. However, a large number of CQs is required to satisfy the asymptotic distributional assumptions for $Cal(s_i)$ in case of DRVs [194] – more than in case of CRVs. Since surgeons were asked to complete the SEJ in addition to the vignette study, a high risk of response fatigue was anticipated. Hence, in consultation with an expert in the field of SEJs, it was decided to elicit CRVs in the CQs, despite the target variables being discrete event probabilities. Based on a rule of thumb, the number of required CQs would then reduce from 25 to 14.

A summary of all methodological steps pursued during the SEJ, from preparing the instruments to obtaining the pooled estimates, is shown in Figure 3.4.

3.2.9 Statistical Analysis of Surgeon Characteristics

Pearson's correlation test was conducted to examine whether surgeons with more years of experience gave more informative 30-day mortality risk estimates during the SEJ. Potential differences in conservative treatment choice proportions between medical specialists and surgical residents were tested with Wilcoxon's rank sum test. For both tests, p-values below 0.05 were considered statistically significant. Finally, to examine the degree to

which treatment recommendations could be explained by surgeons' personal preferences, rather than changes in attribute levels, the intraclass correlation coefficient (ICC) was computed using the approximation described by Sommet and Morselli [247].

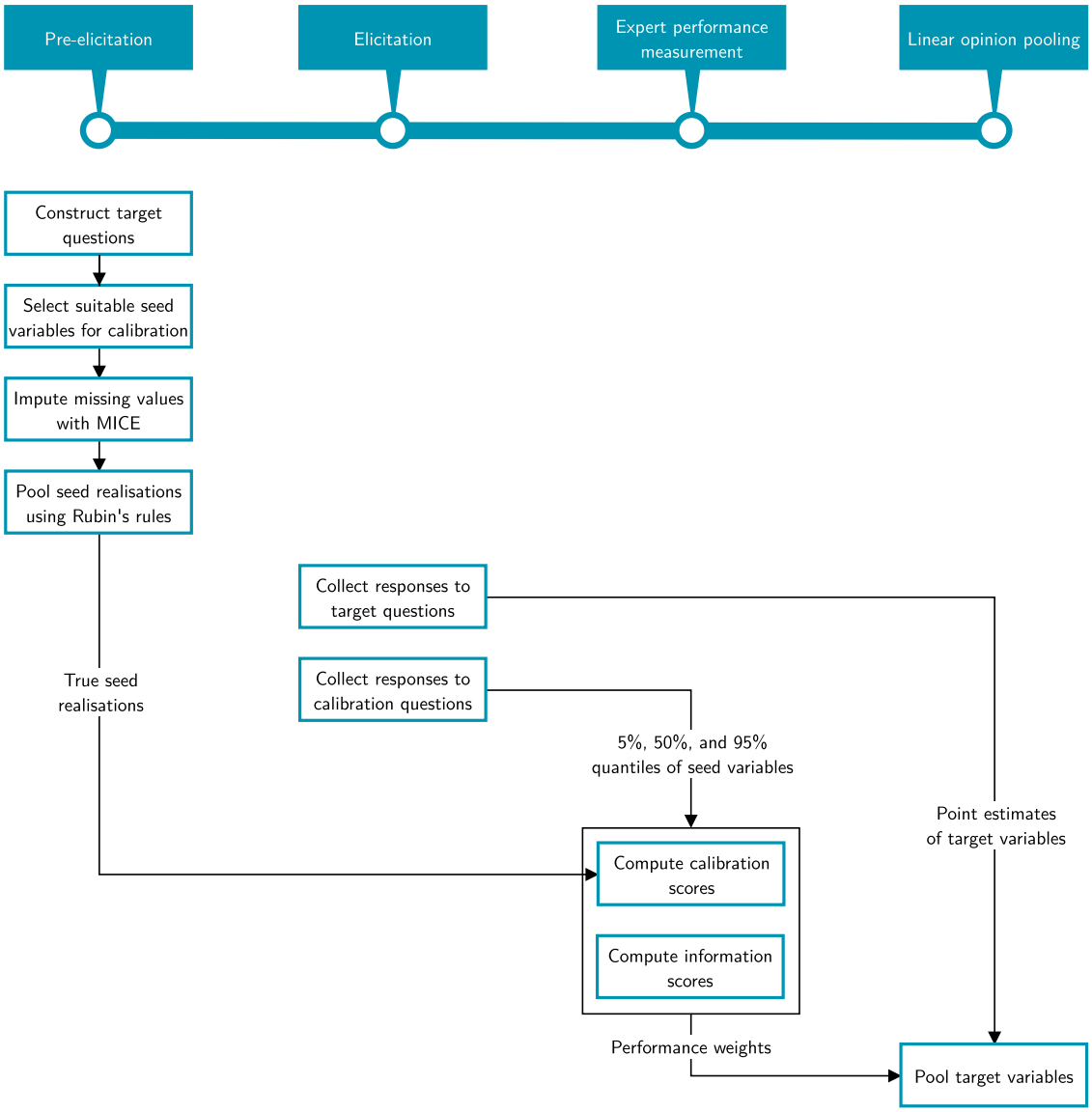

Figure 3.4: Overview of the pursued methodological steps to conduct the structured expert judgement. The steps have been categorised into four phases: pre-elicitation, elicitation, expert performance measurement, and linear opinion pooling.

3.3 Results

3.3.1 Respondents

In total, 21 surgeons were recruited to participate in the vignette study and the SEJ. This resulted in 14 (6 medical specialists and 8 surgical residents) and 9 (4 medical specialists and 5 surgical residents) complete responses respectively. The medians and interquantile

ranges of years of experience for medical specialists and surgical residents were 11.25 (8.50-18.13) and 4.00 (2.75-5.00) respectively.

3.3.2 Power Analysis

Based on the sample of 14 surgeons in the vignette study, the power curves depicted in Figure 3.3 showed that significance tests for gender, fracture type, dementia, age, functional status in ADL, and institutional residence, and ASA classifications attained a power below 60%. Hence, detection of significant effects was only anticipated for end-stage renal failure, metastatic carcinoma, and severe heart failure.

3.3.3 Results from the Vignette Study

Convergence of the Markov chains was confirmed visually by inspection of the trace plots and autocorrelation plots (see Figures B.2-B.5 of Appendix B.3). Furthermore, upon doubling the posterior draws from 15,000 to 30,000 to test for local convergence, Geweke's convergence test yielded p-values above 0.05 for all part-worth utilities. Therefore, for each part-worth utility, it could be assumed that the Markov chains had converged to a stationary distribution. Upon inspecting the frequency histograms of the posterior draws (see Figure B.6 of Appendix B.3), the distributions appeared to be smooth without substantial gaps between the bins. Hence, 15,000 samples were ample enough to adequately represent the posterior distributions.

As shown in Table 3.3, the directions of the effects of the β-coefficients were mostly congruent with prior expectations. Only for fracture type, discrepancies with the a priori hypotheses were observed. Based on the associated surgical procedures, it was anticipated that the perceived benefit of conservative treatment would be the highest for displaced femoral neck fractures, followed by extracapsular fractures and undisplaced femoral neck fractures. The point estimates of the β-coefficients, on the other hand, indicated indifference between displaced and undisplaced femoral neck fractures. Additionally, a slight preference for conservative treatment was observed for undisplaced femoral neck fractures, compared to extracapsular fractures.

Amongst the inspected patient attributes, only four showcased 95% CrIs which did not overlap with the null effect. In descending order of relative importance, these were metastatic carcinoma (18.5%), severe heart failure (17.1%), end-stage renal failure (15.8%), and dementia (14.9%). Collectively, these attributes accounted for 64.1% of the relative importance. From the estimated part-worth utilities, comorbid conditions appeared to increase the perceived benefit of conservative treatment the most amongst surgeons.

All substantive conclusions, i.e. whether the CrIs of the β-coefficients were non-overlapping with the null effect, were robust with respect to decreased informativeness of priors. Additionally, the sensitivity analysis revealed that the informativeness embedded into the priors had little influence on the β-coefficients of end-stage renal failure, preoperative residence, functional status, gender, and age. The prior influence was moderate for severe heart failure, dementia, and fracture type. Finally, priors were highly influential for the effect estimates of metastatic carcinoma and ASA classification.

Table 3.3: Part-worth utilities of patient attributes describing surgeons' preferences for recommending conservative treatment. The 95% credible intervals were estimated by the 2.5% and 97.5% quantiles of the posterior samples. Model estimates are provided for two configurations of the hierarchical Bayesian logit model, i.e. with informative and noninformative priors. Outcome differences for the two configurations are expressed in terms of relative deviations.

Attribute	**Level**	**Estimates based on informative priors**				**Estimates based on noninformative priors**				**Relative deviation**	
		β	**SD**	**OR (95% CrI)**	**Rel. imp.**	β	**SD**	**OR (95% CrI)**	**Rel. imp.**	**OR**	**Rel. imp.**
Metastatic carcinoma[c]	Present	1.495	0.382	4.46 (2.13-9.49)	18.5%	1.870	0.508	6.49 (2.37-17.42)	21.7%	45.5%	17.3%
	Absent*	0.000				0.000					
Severe heart failure[b]	Present	1.381	0.382	3.98 (1.93-8.11)	17.1%	1.522	0.416	4.58 (2.10-10.68)	17.6%	15.1%	2.9%
	Absent*	0.000				0.000					
End-stage renal failure[a]	Present	1.275	0.364	3.58 (1.78-7.28)	15.8%	1.325	0.435	3.76 (1.58-9.06)	15.4%	5.0%	2.5%
	Absent*	0.000				0.000					
Dementia[b]	Present	1.201	0.370	3.32 (1.61-6.83)	14.9%	1.368	0.435	3.93 (1.72-9.63)	15.9%	18.4%	6.7%
	Absent*	0.000				0.000					
Preoperative residence[a]	Institution	0.663	0.377	1.94 (0.93-4.06)	8.2%	0.668	0.431	1.95 (0.84-4.63)	7.7%	0.5%	6.1%
	Home*	0.000				0.000					
ASA classification[c]	ASA IV	0.654	0.413	1.92 (0.87-4.32)	8.1%	0.409	0.465	1.51 (0.63-3.87)	4.7%	21.4%	42.0%
	ASA III*	0.000				0.000					
Functional status[a]	Severe handicaps	0.565	0.358	1.76 (0.88-3.54)	7.0%	0.524	0.404	1.69 (0.76-3.81)	6.1%	4.0%	12.9%
	No severe handicaps*	0.000				0.000					
Gender[a]	Male	0.464	0.384	1.59 (0.75-3.30)	5.7%	0.450	0.415	1.57 (0.70-3.56)	5.2%	1.3%	8.8%
	Female*	0.000				0.000					
Fracture type[b]	Extracapsular fracture	-0.199	0.441	0.82 (0.35-1.97)	2.5%	-0.388	0.500	0.68 (0.25-1.81)	4.5%	17.1%	80.0%
	DFN fracture	0.005	0.464	1.00 (0.41-2.53)		-0.147	0.500	0.86 (0.33-2.33)		14.0%	
	UFN fracture*	0.000				0.000					
Age[a]	$\geq$ 90 years	0.184	0.384	1.20 (0.56-2.49)	2.2%	0.105	0.404	1.11 (0.50-2.44)	1.2%	7.5%	45.5%
	80-89 years*	0.000				0.000					

β part-worth utility coefficient, *SD* standard deviation of posterior draws, *OR* odds ratio, *CrI* credible interval, *Rel. imp.* relative importance, *ASA* American Society of Anaesthesiologists, *DFN* displaced femoral neck, UFN undisplaced femoral neck

The relative deviation was computed as 100 $\times$ |(model with informative prior)-(model with noninformative prior)|/(model with informative prior)

* Reference level

[a] Minimally influenced by prior specification

[b] Moderately influenced by prior specification

[c] Highly influenced by prior specification

3.3.4 Results from the Structured Expert Judgement

Expert Performance

Table 3.4 depicts the calibration scores, information scores, and performance-based weights of the 9 surgeons who completed the SEJ. It was noteworthy that the surgical residents obtained higher calibration scores than the medical specialists. Moreover, the surgical residents were the only ones to achieve calibration scores corresponding to p-values above 0.05, indicating that they were well-calibrated.

Overall, the information scores appeared similar across surgical residents and medical specialists. However, across all surgeons, the information scores differed by a factor of 3 at most. This indicated substantial differences in the degrees of certainty expressed in surgeons' assessments of the CQs. It was noteworthy that the varying degrees of certainty could not directly be explained by how experienced surgeons were. Pearson's correlation test revealed an insignificant negative correlation between information scores and years of experience ($r = -0.139$, $p = 0.722$).

Table 3.4: Overview of calibration scores, information scores, performance-based weights.

Occupation	Experience	Calibration score	Information score	Performance weight
Medical specialist	8 years	3.24×10^{-7}	2.11	8.60×10^{-7}
Medical specialist	10 years	0.01	1.70	0.03
Surgical resident	4 years	0.53	0.71	0.47
Surgical resident	5 years	0.32	1.15	0.46
Surgical resident	5 years	9.58×10^{-7}	2.18	2.62×10^{-6}
Surgical resident	0.5 years	9.17×10^{-6}	2.07	2.39×10^{-5}
Medical specialist	20 years	3.90×10^{-12}	0.70	3.42×10^{-12}
Surgical resident	2 years	0.04	0.66	3.66×10^{-2}
Medical specialist	20 years	1.33×10^{-15}	1.59	2.65×10^{-15}

Upon inspecting the performance-based weights, it became apparent that 93% of the cumulative weight in the linear opinion pool was accounted for by the two surgical residents with the highest calibration scores of 0.53 and 0.32. Although the calibration scores differed by a factor of 1.65, the performance-based weights were nearly equal (0.47 vs 0.46). The reason for this was that the most statistically accurate surgeon specified wider 90% CrIs, leading to a lower information score.

Pooled Estimates

An overview of the 30-day mortality probability estimates obtained through linear opinion pooling with equal weights (EWs) and performance-based weights (PWs) is shown in Table 3.5. For each vignette, the PW estimates were consistently lower than the EW estimates. The pooled probabilities across all vignettes ranged between 20.7-62.7% and 11.9-50.8% for EW and PW respectively. The differences were most pronounced for vignette 1 (EW: 20.7%, PW: 11.9%), vignette 11 (EW: 43.6%, PW: 26.8%), vignette 13 (EW: 37.1%, PW: 21.9%), and vignette 16 (EW: 38.6%, PW: 26.8%).

3.3.5 Individual Risk Perceptions and Treatment Preferences

Figure 3.5 depicts the relationship between surgeons' individual 30-day mortality risk perceptions, and the probability that they would recommend conservative treatments. For each vignette, the part-worth utilities and the surgeon-specific intercepts were used to predict the probabilities of conservative treatment recommendations.

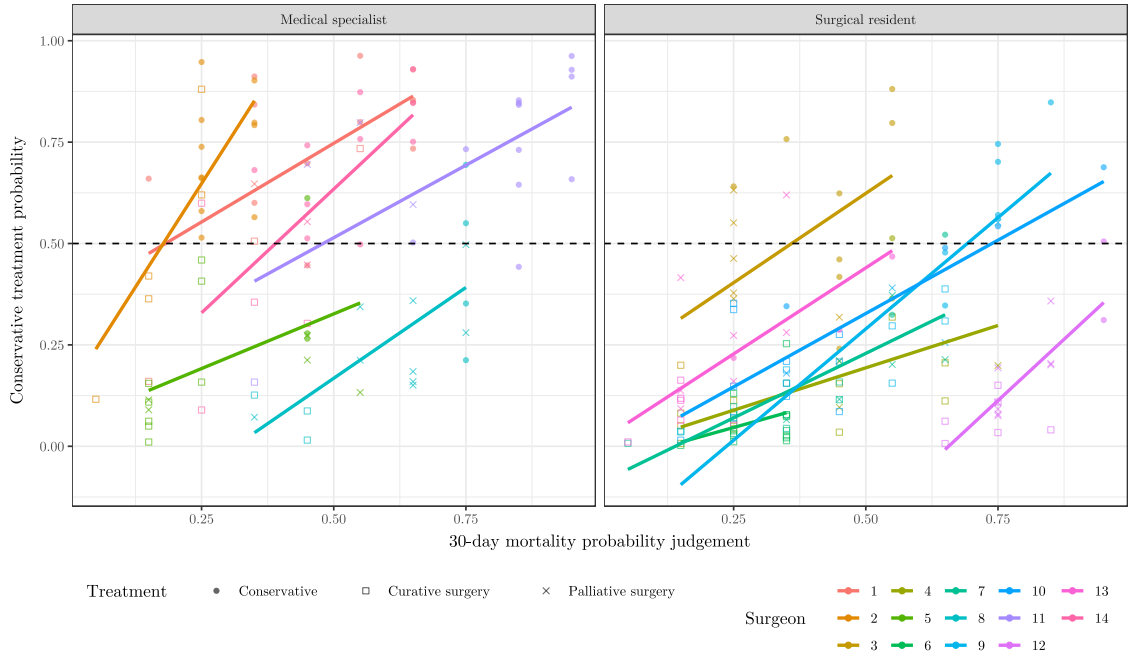

Figure 3.5: Relationships between subjective mortality risk perceptions and inclination to recommend conservative treatments per surgeon.

Based on these results, heterogeneity was identified across two dimensions. Firstly, even though all surgeons examined the same vignettes, their perceptions of 30-day mortality risks differed substantially. For surgical resident 6, for instance, the 30-day mortality bandwidth across all 16 vignettes was confined to 15-35%. This entailed that changes in attribute levels minimally influenced the surgeon's risk perceptions. Conversely, for surgical resident 10, the 30-day mortality bandwidth spanned a much wider range of 15-95%. This indicated that the surgeon's risk perceptions were sensitive to changes in attribute levels.

Secondly, preferences for conservative treatment appeared to differ considerably from surgeon to surgeon. The ICC was estimated at 0.299. This meant that 29.9% of the treatment preferences was explained by personal differences between surgeons. Overall, medical specialists appeared to be more inclined to prefer conservative treatment over operative treatment than surgical residents. According to Wilcoxon's rank sum test, the difference in conservative treatment choice proportions was statistically significant ($p = 0.046$). To exemplify this difference, consider medical specialists 1, 2, 11, and 14, who showcased conservative treatment choice proportions of 11/16, 12/16, 12/16, and 10/16 respectively. The surgical residents generally showcased lower choice proportions, with

one resident even preferring operative treatment with curative intentions over conservative treatment for all 16 vignettes. Therefore, practice variation appeared to be substantial.

Although heterogeneity was observed in risk perceptions and treatment preferences, the trend lines shown in Figure 3.5 indicated that preferences for conservative treatment generally increased with the perceived 30-day mortality risk. However, there was no universal 30-day mortality probability threshold which could serve as a recommended tipping point for favouring conservative treatment. To substantiate this claim, consider the choice behaviours of medical specialist 2 and surgical resident 12 in Figure 3.5. The former showcased a conservative treatment choice proportion of 12/16 across a 30-day mortality probability bandwidth of 5-35%, whereas the latter showcased a conservative treatment choice proportion of 2/16 across a 30-day mortality probability bandwidth of 65-95%. This entailed that the 30-day mortality risk tipping point at which the benefit of conservative treatment was acknowledged differed considerably between the surgeons.

3.3.6 Uncertainty in Treatment Recommendations

For each vignette, surgeons were asked to express how certain they were about the optimality of their treatment recommendation. The results are summarised in Figure 3.6. The greatest degree of certainty was expressed for vignettes 1 and 4. Based on the a priori assumed effect sizes of the respective attribute levels, these vignettes described the theoretically best and theoretically worst candidates for surgery respectively. As shown in Table 3.5, all 14 surgeons agreed that operative management with curative intentions was the best choice for vignette 1. For vignette 4, 12/14 surgeons agreed that conservative treatment was the best choice.

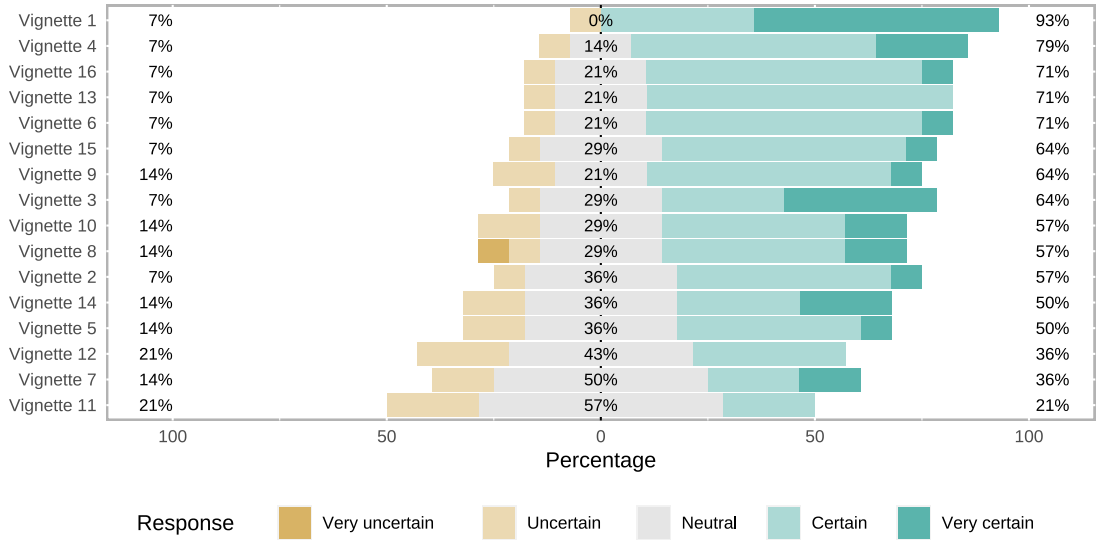

Figure 3.6: Summary of likert scale data depicting the certainty expressed in the optimality of treatment recommendations for each vignette.

Table 3.5: Overview of descriptive statistics for each of the vignettes. The choice proportions for conservative treatments, surgery with palliative intentions, and surgery with curative intentions were reported relative to the total number of respondents in the vignette study. The 30-day mortality probabilities estimated by surgeons were summarised with equally weighted pooled estimates, performance weighted pooled estimates, and the interquartile range of the sample.

ID	ASA	Fracture	Gender	Age	Functional status	Residence	Dementia	SHF	Metastasis	ESRF	Choice proportions			30-day mortality %		
											CTP	PSP	CSP	EW	PW	IQR
1	III	DFN	Female	80-89	No severe handicaps	Home	Absent	Absent	Absent	Absent	0/14	0/14	14/14	20.7	11.9	(15-25)
2	IV	UFN	Female	90+	No severe handicaps	Home	Absent	Present	Absent	Present	4/14	3/14	10/14	48.6	40.8	(45-65)
3	IV	DFN	Female	90+	No severe handicaps	Institution	Present	Absent	Present	Present	10/14	2/14	2/14	60.0	46.1	(55-75)
4	IV	EXT	Male	80-89	Severe handicaps	Institution	Absent	Present	Present	Present	12/14	1/14	1/14	62.9	50.8	(55-85)
5	IV	EXT	Female	90+	Severe handicaps	Home	Absent	Absent	Present	Absent	2/14	7/14	5/14	38.6	35.8	(25-45)
6	III	DFN	Male	80-89	Severe handicaps	Home	Present	Absent	Absent	Present	4/14	3/14	7/14	42.1	36.1	(25-65)
7	IV	UFN	Female	80-89	Severe handicaps	Institution	Present	Absent	Absent	Present	6/14	6/14	2/14	50.7	45.7	(35-75)
8	IV	UFN	Male	80-89	No severe handicaps	Home	Present	Present	Present	Absent	8/14	2/14	4/14	52.9	41.5	(35-75)
9	IV	DFN	Male	90+	Severe handicaps	Home	Present	Present	Absent	Absent	8/14	4/14	2/14	52.9	50.4	(45-65)
10	III	EXT	Female	90+	Severe handicaps	Home	Present	Present	Absent	Present	7/14	5/14	2/14	55.0	41.5	(35-75)
11	IV	DFN	Female	80-89	Severe handicaps	Institution	Absent	Present	Absent	Absent	3/14	8/14	3/14	43.6	26.8	(35-55)
12	III	DFN	Male	90+	No severe handicaps	Institution	Absent	Present	Absent	Present	5/14	4/14	5/14	45.7	36.1	(35-55)
13	III	UFN	Male	90+	Severe handicaps	Institution	Absent	Absent	Absent	Absent	2/14	1/14	11/14	37.1	21.9	(15-45)
14	IV	EXT	Male	80-89	No severe handicaps	Home	Absent	Absent	Absent	Present	2/14	0/14	12/14	37.9	36.2	(25-45)
15	IV	EXT	Male	90+	No severe handicaps	Institution	Present	Absent	Absent	Absent	3/14	4/14	7/14	40.7	31.8	(25-55)
16	III	EXT	Female	80-89	No severe handicaps	Institution	Present	Present	Absent	Absent	4/14	4/14	6/14	38.6	26.8	(15-65)

ASA American Society of Anaesthesiologists physical status classification, *SHF* severe heart failure, *ESRF* End-stage renal failure, *CTP* conservative treatment proportion, *PSP* palliative surgery proportion, *CSP* curative surgery proportion, *EW* equally weighted pooled estimate, *PW* performance weighted pooled estimate, *IQR* interquartile range, *DFN* displaced femoral neck, *UFN* undisplaced femoral neck, *EXT* extracapsular

However, a large proportion of treatment proposals for which certainty in optimality was expressed did not necessarily imply a consensus between surgeons. Although 10/14 surgeons expressed to be certain, or even very certain, about the optimality of their treatment recommendations for vignette 16, the treatment choice proportions were rather divided. In total, 6/14 surgeons recommended conservative treatment, 4/14 surgeons recommended operative treatment with curative intentions, and 4/14 surgeons recommended operative treatment with palliative intentions. It was noteworthy that vignette 16 also exhibited the greatest degree of disagreement amongst surgeons with respect to the estimated survival prognosis. As depicted in Table 3.5, the 30-day mortality probability assessments of vignette 16 showcased the widest interquartile range of 15-65%.

Surgeons expressed the greatest degree of uncertainty about their treatment proposals for vignettes 7, 11, and 12. Vignettes 7 and 11 described ASA IV patients with at most one known physical comorbidity and at least two indicators of poor cognitive or functional status. Vignette 12 described an ASA III patient with two known physical comorbidities and a single indicator of poor cognitive or functional status. What these vignettes have in common, is that they describe patients with a moderately poor health status and pre-fracture hindrances in activities of daily living. As outlined previously, indicators of poor health status had the highest relative importance in preoperative decision-making. This indicated that medical grounds supported surgeons' decisiveness the most. It is postulated that patients with moderately poor health status, rather than a severely poor health status, induced a greater degree of decision uncertainty due to a paucity of medical grounds to base the treatment recommendation on.

Since the vignettes represented a simplification of reality, it was possible that surgeons missed certain nuances that could have helped them to assess the patient cases more confidently. In total, nine surgeons provided feedback on what information they missed in the vignette descriptions. Two surgeons expressed that they did not need any additional information. Amongst the remaining surgeons, the wish for additional information concerned two primary themes. The first theme was conceptualised as: *a need for an enriched medical profile*. In particular, surgeons were interested in knowing patients' pulmonary status, their survival prognosis for metastatic cancer, the motivation behind high ASA scores, and patients' mobility status. The second theme was conceptualised as: *decision-making as a socially malleable process*. Surgeons expressed that they missed the social nuances of being able to look patients in the eye and asking them and their relatives about their personal treatment preferences. Additionally, one surgeon expressed that second opinions from geriatricians, anaesthesiologists, and cardiologists could have helped in shaping a better treatment proposal.

3.4 Discussion

This paper reports on the results of the first health preference study examining surgeons' perceived benefit of conservative treatment for frail geriatric hip fracture patients. The results showed that metastatic carcinoma, severe heart failure, end-stage renal failure, and dementia had the strongest influence on surgeons' preferences to recommend conservative treatment. Amongst the 10 examined attributes, these were the only ones with 95% CrIs which did not overlap with the null effect. Furthermore, it was observed that

the probability of preferring conservative treatment over operative treatment, generally increased with the mortality risks prognosticated by surgeons. These findings underline and confirm that comorbidities leading to increased mortality risk are some of the strongest indicators of decreased benefit of operative treatment [21, 26].

However, some of these findings were unexpected given the a priori hypotheses. Firstly, based on the power analysis with an assumed OR of 1.4, no significant effect was expected to be found for the influence of dementia. In fact, with an observed OR of 3.32, dementia appeared to have a subtantially higher influence on the perceived benefit of conservative treatment than hypothesised. Secondly, the estimated effect size of metastatic carcinoma appeared to be highly influenced by the specification of the informative prior. Similar to before, the a priori assumed OR of 2.5 was substantially smaller than the observed ORs of 4.46 (2.13-9.49) and 6.49 (2.37-17.42) for informative and noninformative priors respectively. This gives rise to the question whether the influence of these attributes was undervalued in the a priori hypotheses, or whether surgeons overvalued the utility of conservative treatment for these attributes.

In retrospect, we would like to plead for the former. The a priori assumed effect sizes of the attributes were solely estimated based on their prognostic value for 30-day mortality. Initially, the assumption was made that 30-mortality risk could function as a viable proxy to model the β-coefficients in the vignette study, since risk of early death is a leading argument to elect conservative treatment according to the national guidelines [26]. However, 30-day mortality risk alone may not be sufficient to fully encompass the utility of conservative treatment, as it overlooks QoL considerations [28]. Spronk et al. [21] found that over 90% of the 271 surveyed healthcare providers expressed that a poor postoperative QoL prospect was a common reason for them to treat frail geriatric hip fracture patients conservatively. Hence, as QoL was not accounted for in the a priori estimates, we may have undervalued the influence of dementia and metastatic carcinoma. To further substantiate these claims, important QoL considerations for both conditions will be delineated.

Firstly, it is increasingly acknowledged that dementia is a terminal condition [248–250] which necessitates palliative care assessments [251]. This necessity is particularly pronounced in advanced stages of dementia with inclinations towards self-neglect, e.g. in the form of malnutrition due to dysphagia [248]. In end-of-life care for demented older adults, Dutch clinicians agree that forgoing artificial nutrition and hydration (ANH) could be good medical practice [252], as ANH prolongs patients' lives at the expense of serious discomfort [253]. These findings demonstrate that improving the QoL of demented patients may in fact entail improving their quality of dying. However, these circumstances may not be applicable to all demented hip fracture patients, but primarily to those with advanced dementia [189]. Nevertheless, since preoperative dementia is a well-known significant risk factor for postoperative delirium, surgery may accelerate patients' cognitive decline [254–256]. With these outcomes in mind, the utility of conservative treatment may come from poor postoperative QoL prognoses, on top of increased mortality risk.

Secondly, recovery-oriented surgery is unlikely to improve the well-being of geriatric hip fracture patients who are debilitated by advanced malignancy [24]. While pain reduction could be a viable motive to elect surgery, the treatment's benefit depends on the patient's

age and health status. Preference studies have shown that cancer patients aged 65 years and older are less willing to trade prolonged survival for decreased QoL than their younger counterparts [257]. Concordantly, aggressive oncological treatments occur less frequently in geriatric patients [258]. Especially for those who are frail and suffer from metastasis, best supportive care could be preferred due to its acceptable outcomes with respect to QoL [259–261]. Therefore, considering the implications of frailty and patients' end-of-life preferences, QoL aspects may have contributed to the perceived utility of conservative treatment for hip fracture patients with metastatic cancer.

While several patient attributes were found to be critical for preoperative decision-making, it should be noted that surgeons' treatment preferences were rather heterogeneous. In fact, personal differences between surgeons accounted for 29.9% of the explained variance of surgeons' treatment recommendations. These results suggest that practice variation may be substantial. On the one hand, heterogeneity in stated preferences could be attributable to the simplified nature of the vignettes, leading to a lack of nuances which could have helped surgeons assess the patient cases more confidently and reliably. On the other hand, even for vignettes where surgeons consistently expressed (high) certainty for the optimality of their treatment recommendations, stated preferences remained divided. These observations are most likely reflecting the lack of guidelines on electing conservative treatment as a palliative treatment option.

Besides that, substantial heterogeneity in 30-day mortality risk perceptions was observed as well. In the most extreme case, one surgeon estimated 30-day mortality risks to range between 15-35% across all vignettes, whereas another surgeon provided an estimation range of 65-95%. This exemplifies the need for objective 30-day mortality prediction models to streamline risk perceptions across clinicians. Through the SEJ, an attempt was done at forging a rational consensus between surgeons' dispersed risk estimates. Through linear opinion pooling with performance-based weights, the expert model yielded a 30-day mortality prediction range between 11.9-50.8% across all vignettes. However, the maximum risk estimate appeared to be rather low, considering that it was the prognosis for a male institutionalised ASA IV patient, between the ages of 80-89 years, with severe functional handicaps, severe heart failures, metastatic cancer, and end-stage renal failures.

To validate the expert-driven estimate, a comparison was made with data-driven prediction models. An overview of the maximum predicted risks and the respective predictor variables of the Nottingham Hip Fracture Score (NHFS) [176], Almelo Hip Fracture Score (AHFS) [87], AHFS^{90} [262], and Brabant Hip Fracture Score (BHFS) [83] is shown in Table 3.6. This overview shows that most predictors also appear in the vignettes. The vignettes, however, include three strong predictors for 30-day mortality that are not included in the prediction models: metastatic cancer, severe heart failure, and end-stage renal failure. Based on the systematic review of Chapter 2, it can be concluded that these predictors have larger effect sizes than most of the other predictors considered in the NHFS, AHFS, AHFS^{90}, and BHFS. Yet, the expert model only attained a marginally higher maximum risk than the NHFS and BHFS, and a lower maximum risk than the AHFS and the AHFS^{90}. As the maximum AHFS and AHFS^{90} were computed in a relatively healthy population compared to the vignettes, the expert model is likely downward biased for the most vulnerable patients. Thus, patients who are at high risk of early mortality are potentially underidentified in practice.

Table 3.6: Comparison of objective 30-day mortality risk scores and subjective expert-driven estimates. The attributes included in each of the prediction models is indicated with an X.

Attribute	**Maximum 30-day mortality probabilities**				**Surgeons' judgements**	
	NHFS (45.0%)	AHFS (68.4%)	$AHFS^{90}$ (64.5%)	BHFS (46.6%)	EW (62.9%)	PW (50.8%)
Age	X	X	X	X	X	X
Gender	X	X	X	X	X	X
Preoperative residence	X	X	X	X	X	X
History of any malignancy	X	X		X	X[a]	X[a]
Cognitive impairment	X	X	X		X	X
Admission haemoglobin	X	X	X	X		
ASA classification		X	X		X	X
Number of comorbidities	X	X				
Mobility		X				
COPD				X		
Diabetes				X		
Functional status					X	X
Severe heart failure					X	X
End-stage renal failure					X	X
Fracture type					X	X

NHFS Nottingham Hip Fracture Score, *AHFS* Almelo Hip Fracture Score, $AHFS^{90}$, Almelo Hip Fracture Score in patients aged $\geq$ 90 years, *BHFS* Brabant Hip Fracture Score, *EW* equally weighted pooled estimate, *PW* performance weighted pooled estimate, *ASA* American Society of Anaesthesiologists physical status classification, *COPD* chronic obstructive pulmonary disease

[a] Malignancy was exclusively defined as metastatic cancer

Several limitations may have caused the SEJ to yield poor risk estimates for the most vulnerable patients. The underestimation may have been due to chance, since only nine surgeons completed the SEJ. Hence, replication of the study in a larger cohort is necessary to validate the results. Nevertheless, two exceptionally well-calibrated surgeons with calibration scores of 0.53 and 0.32 were observed in this small sample, accounting for a cumulative weight of 93% in the pooled estimates. Based on the premise of the SEJ, it is counter-intuitive that these surgeons underestimated the 30-day mortality risks for the most vulnerable patients. It is postulated that the CQs did not capture the required range of expertise for the assessments of the diverse vignettes, since the seed variable realisations were merely confined to 30-day mortality rates between 3.9-33.2%. Since the SEJ instrument was calibrated to mortality rates of relatively healthy patients, a high calibration score did not reflect accurate predictions for high-risk patients. This is a challenge for SEJs in this field, since data on more frail hip fracture patients are limited due to the lack of national registrations of severe comorbidities such as metastasis.

Another limitation is that all risk estimates and preferences were elicited, assuming that the judgements of surgeons alone could represent the clinical decision context. However, surgeons reported that patients' preferences and fellow clinicians' opinions could have contributed to better-informed judgements. Hence, it is questionable whether the aforementioned assumption is justified. From literature, it is known that a patient's

personal wish is one of the strongest deciding factors for surgeons to treat patients conservatively [21]. Additionally, consultations with geriatricians may significantly increase the percentage of elected conservative treatments following SDM [189]. Hence, it is evident that decisions in healthcare are shaped by the opinions of multiple stakeholders in the decision context [263]. As these considerations were not accounted for, the external validity of the stated preferences may have been harmed [264]. Therefore, future studies should incorporate the views and values of the core agents in the decision context, to improve the external validity of the stated preferences.

Despite the concerns regarding external validity, we still believe that the contributions of this study are valuable. While it cannot be determined how the stated preferences would have changed if SDM had taken place, SDM should not be taken for granted. Amongst others, lack of time to organise multidisciplinary consultations is a persistent barrier to SDM in management of acute hip fractures [21]. Furthermore, when SDM between patients and surgeons does take place, surgeons' initial judgements may still have a substantial influence due to the power asymmetry in doctor-patient interactions [265]. This is not meant as a criticism, but rather as a reminder that patients put a lot of faith in the expertise and authority of healthcare providers [266]. Based on medical grounds, surgeons may firmly direct towards conservative treatment [267]. Therefore, surgeons' treatment preferences can still offer valuable insights into the utility of conservative treatment for frail geriatric hip fracture patients. However, the results should be interpreted with caution. Most statistical tests were underpowered, meaning that increasing the sample size may lead to the detection of additional significant decision variables.

3.5 Conclusion

This study demonstrated that comorbidities had the strongest influence on surgeons' perceived benefit of conservative treatment. Although surgeons were more inclined to abstain from surgery amongst hip fracture patients for whom they prognosticated higher 30-day mortality risks, heterogeneity in treatment preferences and risk perceptions was substantial. Hence, objective 30-day mortality prediction models should be used in clinical practice to streamline risk perceptions across surgeons. However, objective mortality risk estimates alone are postulated to be insufficient to identify eligible candidates for conservative treatment. Although meta-analyses revealed that some of the examined attributes were of small-to-moderate prognostic value for 30-day mortality, surgeons could still associate them with a high utility of conservative treatment. The increased utility of these attributes is presumably derived from poor postoperative QoL prognoses, in addition to increased 30-day mortality risk. Hence, based on surgeons' stated preferences, more emphasis may need to be put on QoL considerations in the national guidelines, to adequately provide decision support for electing nonoperative management.

CHAPTER 4

A Literature Review of Best Practices in Ambulatory Accelerometry

4.1 Introduction

Over the years, the use of accelerometers in human movement analysis has proliferated amongst clinicians and researchers alike [37, 268–275]. Amongst the applications in healthcare and beyond [275], the impact on rehabilitation care is particularly pronounced [274]. Ambulatory accelerometry provides a promising means to monitor patients' physical activity reliably, continuously, and objectively in an unsupervised manner [268]. Consequently, accelerometers provide rich insights into a patient's physical capacity of being mobile [35, 38], which is beneficial to rehabilitation programmes geared towards restitution of mobility and functional independence [268].

It should be noted that accelerometers do not measure physical activity directly [276]. Most commercially available sensors are triaxial, meaning that they can register accelerations of body segments along the x-, y-, and z-axes. This way, an accelerometer encodes information on both body posture and movement [273, 277–279]. The general premise is that different activities contain distinct information in their accelerations, from which physical activities can be distinguished and quantified using algorithms. Various researchers have successfully applied these principles to develop monitoring systems which can recognise activities of daily living [276, 280–282].

To develop an ambulatory monitoring system, practitioners generally follow the human activity recognition (HAR) chain [283]. The steps along the chain are summarised as follows: (1) data acquisition, (2) preprocessing, (3) segmentation, (4) feature extraction and selection, and (5) classification. In each step, practitioners have to decide which design choices yield the best solution to the task at hand [284]. Although design choices in HAR have been aggregated extensively in reviews [274, 275, 278, 284–287], concrete recommendations for best practices along each step of the HAR chain remain lacking.

To further support clinical researchers with making adequate design choices, this literature review aims to synthesise best practices in ambulatory accelerometry. Evidence synthesis is restricted to HAR paradigms which do not resort to so-called black-box machine learning models, e.g. as described in [288–294], to safeguard the interpretability of the models in clinical implementation processes which may follow. The aim of this review is to provide recommendations for each step along the HAR chain to support the development of ambulatory monitoring systems.

4.2 Data Acquisition

Data acquisition refers to the process of measuring bodily acceleration signals. Important design decisions in this process include determining how many accelerometers should be used for data collection, and where they should be positioned along the body to properly characterise physical activities.

The Number of Accelerometers

The number of required accelerometers depends on the complexity of the to-be-recognised activities. In the field of physical activity monitoring, it is recommended to compartmentalise high-level descriptions of ADL tasks such as *going to bed* into simpler atomic operations: *walking* (to the bedroom), *standing* (next to the bed), performing a *stand-to-sit transfer*, performing a *sit-to-lie transfer*, and finally *lying down* [295]. Such atomic operations can be recognised accurately with fewer accelerometers, and provide sufficient insights into how physically active or how sedentary individuals are [276].

As shown in Table 4.1, most studies already attain satisfactory recognition rates with only one [296–304] or two [305–309] accelerometers for a wide range of different HAR tasks. Studies examining the influence of the number of sensors on HAR performance have shown that using more than two accelerometers only improves the performance marginally [307, 310]. Hence, the design choice regarding the number of accelerometers used for data acquisition in HAR, can generally be simplified to choosing between measurement setups comprising one or two accelerometers.

Two primary considerations to support the aforementioned design choice have been identified in literature. On the one hand, using a single accelerometer could be favoured from a user-centred standpoint. Multiple researchers in the field have argued that the number of body-worn sensors should be minimised to reduce the discomfort of wearing the devices [284, 303, 307, 311]. Especially for applications related to patient monitoring, discomfort could induce compliance issues [307], resulting in a loss of prognostic insights. On the other hand, performance-based considerations could steer practitioners towards favouring two accelerometers. In some cases, a single accelerometer may yield similar data patterns for different activities [296]. Then, the data from a single accelerometer could be insufficiently discriminative between different activities to provide satisfactory recognition rates and a second accelerometer may be necessary [274].

Table 4.1: Summary of accelerometer-based human activity recognition studies.

Study	Activities	No. sensors	Sensor placement	Features	Classifier	Performance
[296]	Walking, standing, sitting, lying down, stand-to-sit transfer, sit-to-stand transfer, lie-to-stand transfer, stand-to-lie transfer	1	Waist	Mean, signal magnitude area, (shifted) delta coefficients	GMM	Accuracy: 0.92
[312]	Walking, sitting, standing, lying down, running, nordic walking, cycling ascending stairs, descending stairs, vacuuming, ironing, rope jumping	3	Chest, wrist, ankle	Variance, skewness, kurtosis, energy, axial correlations, root mean square, mean absolute value, harmonic mean zero crossing rate, Wilson amplitude, slope sign change, cumulative length	KNN	F1-score: 0.97
[283]	Walking, jogging, running, jumping, trunk twist, waist bends, cycling lateral bend, stretching, arm elevations, shoulder rotations, arm rotations, crouching, rowing, elliptical bike	9	Back, lower arms, upper arms, calves, thighs	Mean	KNN	F1-score: 0.95
[305]	Walking, sitting, standing, running, lying down, walking carrying items, stretching, scrubbing, climbing stairs, watching TV, eating or drinking, cycling riding elevator, riding escalator, vacuuming working on computer, strength-training folding laundry, reading brushing teeth	2	Thigh, wrist	Mean, spectral energy, spectral energy, axial correlations	Decision tree	Accuracy: 0.81
[306]	Slow walking, fast walking, running, ascending stairs, descending stairs, dancing	1	Waist	Mean, standard deviation, range, root mean square, axial correlations, average peak frequency variance	MLP	Accuracy: 0.90
[297]	Walking, sleeping, eating or drinking, brushing teeth, dressing, ironing, sweeping, washing dishes, watching TV	1	Wrist	Axial correlation, maximum, minimum, root mean square, maximum norm, differences	SVM	F1-score: 0.95

Continued on next page

Table 4.1 (Continued)

Study	Activities	No. sensors	Sensor placement	Features	Classifier	Performance
[307]	Walking, sitting, standing, lying down, jogging, ascending stairs, descending stairs	2	Lower back, thigh	Mean, average mean over 3 axes, standard deviation, average standard deviation over 3 axes, skewness, average skewness over 3 axes, kurtosis, average kurtosis over 3 axes, energy, average energy over 3 axes, axial correlations	SVM	Accuracy: 0.98
[298]	Walking, standing, sit-to-stand transfer, ascending stairs, descending stairs	1	Chest	Mean, variance, standard deviation, median, range, root mean square, minimum, maximum, power, energy, skewness, kurtosis, interquartile range, mean absolute deviation	Random forest	Accuracy: 0.87
[299]	Sit-to-stand/stand-to-sit transfer, stand-to-kneel-to-stand transfer walking, running, jumping, sitting	1	Waist	Variance, range, spectral energy, spectral entropy	KNN	Accuracy: 0.98
[300]	Walking, standing, sitting, lying down, ascending stairs, descending stairs, jogging, cycling	1	One of the following: chest, upper arm, lower arm, waist, lower leg, upper leg	Principal components explaining 95% of the variance, computed from time and frequency domain features	KNN	Accuracy: 0.96
[301]	Walking, standing, sitting, ascending stairs, descending stairs	1	Thigh	Mean, standard deviation	KNN	Accuracy: 0.99
[302]	Walking, standing and sitting (jointly), cycling, other activities	1	Ankle	Mean, standard deviation, minimum, maximum, spectral power, dominant frequency	SVM	F1-score: 0.92
[308]	Walking, standing, sitting, lying down, squatting, crawling, moving hands	2	Wrist, hip	Mean, variance	HMM	Accuracy: 0.87
[303]	Walking, standing, sitting, lying down prone, lying on the left, lying on the right	1	Waist	Mean, minimum, maximum	Naive Bayes	Accuracy: 0.96

Continued on next page

Table 4.1 (Continued)

Study	Activities	No. sensors	Sensor placement	Features	Classifier	Performance
[313]	Walking, sitting, standing, lying down, cycling, ascending stairs, descending stairs, riding elevator, vacuuming, brushing teeth, cleaning whiteboard, reading, typing, running, watching TV	4	Thigh, wrists, neck	Mean, standard deviation	KNN	Accuracy: 0.91
[309]	Walking, running, ascending stairs, descending stairs, jogging, jumping, hopping on left and right leg	2	Waist, ankle	Mean, standard deviation	KNN	Accuracy: 0.95
[304]	Walking, standing, sit-ups, running, ascending stairs, descending stairs, vacuuming, brushing teeth	1	Pelvis	Mean, standard deviation, spectral energy, axial correlations	Random forest	Accuracy: 0.99

Accelerometer Placement

Apart from deciding how many accelerometers a data acquisition setup should comprise, practitioners should also consider where to position them along the human body. As shown in Table 4.1, some of the most frequently used sensor positions in HAR (beyond ambulatory monitoring) include the waist [283, 296, 299, 300, 303, 306, 309], thigh [283, 301, 305, 307, 313], wrist [297, 305, 308, 312, 313], and ankle [302, 309, 312]. An accelerometer's output varies substantially depending upon its placement, even for the same activity [314, 315]. Based on this observation, various studies have examined the optimal placements of accelerometers for various HAR tasks [303, 307, 308, 311, 316].

However, it remains challenging to determine optimal sensor placements based on these results. Firstly, Atallah et al. [311] provided recommendations for sensor placements for the recognition of vaguely defined physical activities, such as low-level activities, medium-level activities, and high-level activities. These findings are of limited utility to most HAR researchers, who aim to classify more fine-grained physical activities [276, 295]. Secondly, Orha and Oniga [316] postulate that placing the sensors on the right hand and right thigh yields optimal results for the recognition of activities such as standing, sitting, walking, ascending stairs, running, and lying down. However, they do not provide quantitative comparisons with other sensor placements to further substantiate the superiority of their recommendation. In addition to the paucity of substantive evidence, others have found that the classification accuracy of activities similar to those described by Orha and Oniga only differed marginally between differently located sensor pairs [303, 307]. Therefore, gaps remain in the understanding of optimal sensor placements.

On a more general note, some guidelines can be followed to develop well-performing and robust HAR models. For many HAR tasks, measurement of whole-body movement is necessary to characterise physical activities. Waist-worn sensors are adequate for this purpose, as the waist is near the centre of mass of the human body, allowing for a good representation of major motions [278]. Additionally, for the recognition of sedentary behaviours, the waist is also postulated to yield consistent and reliable measurements. Lying postures, for instance, could be challenging to characterise when sensors are placed on the limbs: even if individuals perform the same activity of lying down, their limbs may be oriented differently in free-living conditions. These heterogeneous orientations could further complicate HAR [285]. Positioning sensors at the waist could resolve this issue, as this location is less affected by arbitrary peripheral body motions [303].

However, some activities may involve more subtle motions which cannot be captured by measurement of coarse estimates of whole-body movement [315]. Hence, practitioners should at least reason about which muscle regions are activated upon initiating the activity they aim to classify [311]. For instance, omitting sensor placements near the wrist for the recognition of hand movements is bound to compromise the classification performance [308]. However, for HAR tasks specifically focusing on ambulatory monitoring, various systems with high recognition accuracies between 0.92-0.98 have been developed without using any information from the wrist [296, 299, 307]. In case of single accelerometer measurement setups, others have even proposed that the wrist is not biomechanically suitable for the characterisation of most physical activities [317]. Affirmatively, Janidarmian et al. [300] found that wrist-worn accelerometers were most

sensitive to interpersonal differences in movement patterns and therefore generalised poorly. Conversely, their results indicated that the upper leg position was most robust to interpersonal differences. These findings are compelling and convincing, given that eight different sensor placements were evaluated using 14 aggregated data sets describing over 70 physical activities measured across 228 subjects.

4.3 Preprocessing

Preprocessing refers to the procedure of transforming accelerometer signals and removing artefacts from them prior to feature extraction. In many cases, preprocessing steps were not described [299, 301, 305, 307, 312, 313, 318]. Amongst the available preprocessing descriptions, the decomposition of accelerometer signals into DC and AC components through the use of high-pass filters was a common transformation [306, 309]. This separation allows researchers to analyse postural and kinetic information in isolation, through examination of the DC and AC components respectively [273]. The most common application of artefact removal was high-frequency noise suppression e.g. through median filters [296, 303], (weighted) moving average filters [297], and low-pass filters [302]. In some cases, researchers explicitly decided against applying filters for the purpose of noise suppression [283, 298, 304, 319].

In principle, negligence towards artefact removal can be justified. The reason for this is that accelerometer signals are particularly susceptible to information losses if the physical activities of interest are diverse [283]. Most human activities are confined to a narrow and low frequency band, which could be affected by noise suppression filters. The spectral analysis of Sun and Hill [320] demonstrated that the major energy bands of various activities of daily living varied between 0.3-3.5 Hz. They observed that activities with frequencies above 10 Hz had seldom occurrences. Concordantly, others have found that human movements produce accelerations with most of the energy below 15 Hz [321].

4.4 Segmentation

Segmentation refers to the process of dividing physical activity signals into a finite set of smaller activity windows. Three windowing techniques are commonly used, i.e. sliding windows, event-defined windows, and activity-defined windows [286]. The majority of the HAR studies resort to sliding windows, since their low computational costs make them suitable for real-time applications [309]. In this approach, signals are split into intervals of fixed length without inter-window gaps. Upon defining sliding windows, practitioners face two design choices: choosing the length of the window, and deciding whether or not the windows should overlap.

Sliding Window Size

In most studies, relatively short sliding windows of 1-2 seconds provided satisfactory results for the recognition of a wide range of ambulatory and sedentary behaviours [283, 301, 303, 304, 306, 308, 313, 318, 322]. Larger window sizes were also found in literature, ranging from 4-10 seconds [297–299, 302, 304, 305, 307, 312, 323]. As shown in Table

4.1, the studies resorting to small windows and those resorting to large windows aimed to classify similar types of physical activities. Nevertheless, the chosen window sizes differed substantially.

Practice variations could be attributable to a trade-off between classification speed and accuracy. On the one hand, small windows allow computations to be performed quickly, which is particularly desirable in real-time applications [321]. On the other hand, large windows are believed to more informative as they capture physical activities more comprehensively. Banos et al. [283] systematically evaluated the impact of window sizes on the aforementioned trade-off and found that windows of 1-2 seconds yielded the optimal result. Moreover, others have found that increasing the window size beyond 2 seconds could even harm the overall performance of HAR systems [318, 324]. Choosing the window size to be too large could cause multiple different activities to be captured in a single window. This is undesirable in HAR [311], as it introduces ambiguities in the characterisation of specific activities.

Degree of Overlap in Sliding Windows

In most cases, researchers used overlapping sliding windows (OSWs) with 50% overlap [297–299, 301, 303–307, 309, 323]. In some cases, different degrees of overlap of e.g. 20% were used [312, 325]. Amongst the studies identified in this review, reports of non-overlapping windows (NOSWs) were relatively infrequent [283, 302, 308, 313].

There is a popular belief that OSWs improve the performance of HAR systems at the expense of increased computation time. Although evidence in favour of this claim can be found in literature [318], scepticists have argued that OSWs may cause recognition performances to be inflated, especially when the error rate is estimated through activity-stratified cross-validation (CV) where data of the same subject is present in both the training and testing folds [302, 326]. Affirmatively, Dehghani et al. [319] have demonstrated that OSWs could unrightfully inflate the F1-score by 16-21% if the repeated measures structure of HAR data is not accounted for in CV. Upon comparing the classification performance between OSWs and NOSWs through leave-one-subject-out CV, the ostensible superiority of OSWs disappeared and both windows yielded similar generalisation errors. Therefore, practitioners may rightfully favour NOSWs over OSWs, and should even do so if leave-one-subject-out CV is not feasible.

4.5 Feature Extraction and Selection

Feature extraction refers to the process of computing informative attributes from the data enclosed in activity windows. Most features are computed separately for each signal measured by a single accelerometer sensing axis [274]. When multiple triaxial accelerometers are used for HAR, the number of features under consideration could grow substantially. It is unlikely that all features are (equally) informative for HAR tasks. Hence, a feature subset with high discriminative power is generally selected to reduce noise and thereby improve classification performance [286]. This process is known as feature selection.

Feature Extraction

In the application area of HAR, feature quality is postulated to be one of the primary determinants of a high classifier performance [288]. Hence, domain-specific knowledge is generally used to craft informative features [327]. An overview of the most commonly extracted features in HAR is shown in Table 4.2.

Table 4.2: Overview of the most commonly used features in human activity recognition.

Feature	Interpretation	References
Mean / median	The mean provides postural information. It can be used to distinguish between different sedentary behaviours. In case of outliers, the median may be preferred over the mean.	[276, 277, 283, 296, 298, 300–309, 313, 328–331]
Variance / standard deviation / mean absolute deviation	These metrics provides information about the intensity of a physical activity. They can be used to distinguish between sedentary behaviours and physical activities. In case of outliers, the mean absolute deviation may be preferred over the standard deviation.	[277, 298–301, 306–309, 312, 313, 328, 329, 331]
Root mean square	The root mean square is identical to the standard deviation for signals with a mean of zero.	[297, 298, 300, 306, 330, 331]
Axial correlation	Axial correlations can be used to distinguish between activities which involve unidirectional and multidirectional translations. Walking and running, for instance, primarily involve unidirectional translations (anteroposterior), whereas stair climbing involves prominent multidirectional translations (anteroposterior and vertical).	[297, 300, 304–307, 312, 329, 331]
Minimum	The minimum describes the lowest value measured along a single sensing axis of an accelerometer.	[297, 298, 300, 302, 303]
Maximum	The maximum describes the highest value measured along a single sensing axis of an accelerometer.	[297, 298, 300, 302, 303]
Range / interquartile range	The range is the difference between the minimum and the maximum acceleration value. It can be used to distinguish between sedentary behaviours and physical activities. In case of outliers, the interquartile range may be preferred. The latter is the difference between the 25% and 75% quantiles of the acceleration values.	[299, 300, 306, 331]
Skewness	The skewness provides information about the shape of the probability distribution of movement accelerations. It can take on negative and positive values. If the values of the acceleration signals are symmetrically distributed, the skewness is equal to zero. If high values are more likely to occur than low values, the skewness is positive-valued.	[298, 300, 307, 328, 329, 331]
Kurtosis	The kurtosis provides information about the shape of the probability distribution of movement accelerations. It describes how likely extreme values occur in a movement signal.	[298, 300, 307, 328, 329, 331]
Signal magnitude area	The signal magnitude area is a predictor of energy expenditure. It can be used to distinguish sedentary behaviours from physical activities.	[276, 279, 296, 300, 314, 331–333]
Mean-crossing rate	The mean-crossing rate is correlated with the frequency of an activity. It can be used to distinguish between dynamic physical activities.	[283, 312, 329, 331]
Spectral energy	The spectral energy is capable of detecting periodicity in data and it can be used to distinguish between sedentary behaviours and physical activities.	[299, 304, 305, 334, 335]
Spectral entropy	The spectral entropy is used to distinguish between activities with similar spectral energy values such as cycling and running The former corresponds to a nearly uniform motion pattern with a single dominant frequency component, whereas the latter has multiple major harmonics which contribute to an overall higher spectral entropy.	[174, 299, 305, 309]
Sum of squares of wavelet coefficients	The sum of squares of wavelet coefficients can be used to distinguish between different walking patterns (flat surface walking, ascending, and descending).	[323, 336]
Root mean square of wavelet coefficients	The root mean square of wavelet coefficients can be used to distinguish between different walking patterns (flat surface walking, ascending, and descending).	[323, 337]

In general, most features used in HAR can be categorised as either time or frequency domain features (see Figure 4.1). Amongst the formalisms for frequency domain

transformations, a distinction is made between Fourier transforms [338] and wavelet transforms [339]. The latter are more computationally costly, and have been introduced to pose less stringent constraints on the assumption that physical activity signals are stationary processes. Unlike wavelet features, the informativeness of simple time and frequency domain features relies heavily upon this assumption [340]. However, previous HAR studies have found that accounting for nonstationarity in HAR is not necessary, as wavelet features did not yield predictive advantages over the conventional features [302, 309]. Hence, it is postulated that practitioners can safely restrict the candidate features for their HAR system to simple time and frequency domain features.

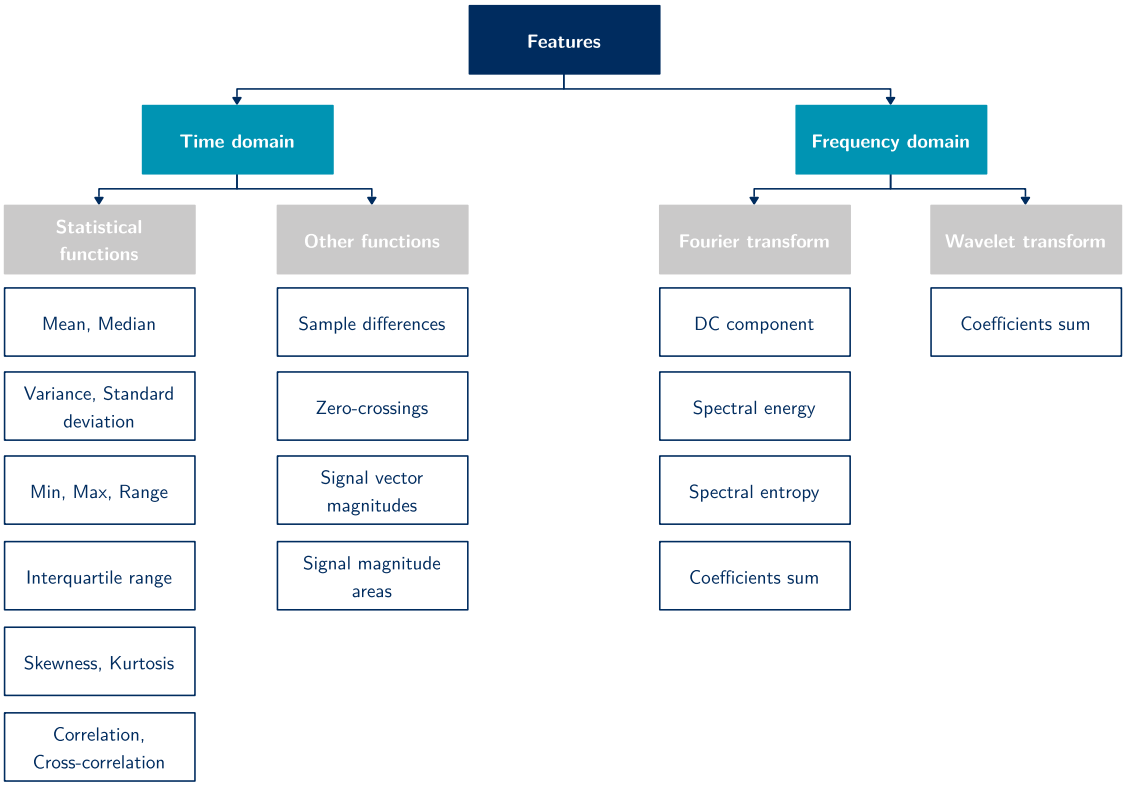

Figure 4.1: Classification of time domain and frequency domain features, adapted from [341, p. 647].

Feature Selection

An overview of the most commonly used feature selection methods in HAR is shown in Table 4.3. It is noteworthy that many studies resorted to manual feature selection based on choices of previous studies [283, 296–298, 301, 302, 304, 306, 307, 309, 342, 343], despite the fact that studies often differ substantially in their experimental design. Amongst the algorithmic feature selection procedures, Relief-F appeared to be most frequently used [299, 303, 311, 344]. The strength of Relief-F lies in its ability to not only select features based on their individual discriminative properties, but also the interactions between them [345]. This is an advantage over solely performing feature selection based on domain knowledge, since the current understanding of how features contribute to HAR (see Table 4.2) is lacking with respect to interactions. Failing to identify interactions may lead to underexploitation of synergy between features.

Table 4.3: Overview of feature selection methods used in human activity recognition studies.

Feature selection mechanism	Description	Studies
Manual feature selection	Selecting features based on the results of previously published studies.	[283, 296–298, 301, 302, 304, 306, 307, 309, 342, 343]
Relief-F	Relief-F uses instance-based learning to select features which have homogeneous values within the same activity class, and heterogeneous values across different activity classes. It is one of the few multivariate feature selection algorithms which is capable of capturing two-way interaction effects between features. However, it is unable to identify redundancy in feature sets.	[299, 303, 311, 344]
Forward-backward selection	In forward-backward selection, features are sequentially added and removed during the model training process, depending upon whether the examined features improve classification performance.	[299, 313]
Principal component analysis	Principal component analysis produces an eigendecomposition of the features' covariance matrix. Each eigenvector describes a new basis vector which can be used for a vector projection to linearly decorrelate features: features can no longer be interpreted on their original scales. By selecting a subset of eigenvectors with the largest eigenvalues, i.e. with the highest explained variance, feature redundancy can be reduced.	[300, 327]
L1 regularisation	L1 regularisation first considers the full feature set and shrinks the influence of irrelevant features to zero during the model training process, by optimising a two-part objective function with a goodness-of-fit term and a penalty for including a large number of features.	[317, 329]
Correlation-based feature selection	Correlation-based feature selection favours features which have high correlations with the activity class, and are minimally correlated with other features. While it is capable of detecting interaction effects between features, it tends to include redundant features as well.	[297, 344]
Fast correlation-based filter	The fast correlation-based filter (FCBF) uses symmetrical uncertainty (SU) as a goodness-of-fit measure, which is defined as the ratio between the information gain and entropy of two features. Features with a high SU are selected by FCBF. While FBCF effectively removes redundant features, it is unable to detect interaction effects between features.	[344]
Minimum redundancy maximum relevance	Minimum redundancy maximum relevance selects features which exhibit a high mutual information with the activity classes (maximum relevance), and a low mutual information between features (minimum redundancy). However, it is unable to identify interactions.	[311]

4.6 Classification

Classification refers to the process of associating a set of features computed from a single window with an activity class. A wide range of classification principles has been used in HAR studies. Supervised machine learning algorithms were most commonly used, which included k-nearest neighbours (KNN) [283, 298, 299, 301, 305, 309, 311–313], decision trees [283, 297, 301, 303, 305, 307], random forests [298, 304, 306, 346], naive Bayes [283, 299, 301, 303, 305, 307, 344], Bayesian networks [301, 303], support vector machines (SVM) [298, 301–303, 306, 307, 344], and multilayer perceptrons (MLP) [297, 303, 306, 307, 313]. A minority of the HAR studies used unsupervised machine learning algorithms such as Hidden Markov Models (HMM) [308] and Gaussian mixture models (GMM) [296]. Apart from classifiers based on machine learning, several studies resorted to rule-based heuristic systems [276, 280, 282, 342, 347, 348].

The choice between rule-based heuristic systems and machine learning models generally boils down to a trade-off between model interpretability and classification performance. While rule-based heuristic systems are easier to interpret and more accessible in terms of parameter tuning [276], machine learning algorithms tend to be superior in terms of classification performance [296]. The current trends in HAR indicate that most researchers favour the latter.

However, for practitioners who want to resort to machine learning, it may remain challenging to choose the best-suited algorithm from the wide range of available options. Various machine learning models have been evaluated simultaneously to identify the best algorithm. Amongst these evaluations, KNN was frequently found to be superior [299–301, 313]. Compared to naive Bayes, decision trees, and a nearest centroid classifier, Banos et al. [283] found that KNN attained the highest F1-score across different feature sets and different window sizes. Moreover, they found that KNN required the least information, i.e. fewer features and smaller window sizes, to attain a high classification performance. Compared to decision trees, discriminant analysis, SVMs, ensemble methods, naive Bayes, and MLPs, Janidarmian et al. [300] found that KNN performed well and retained stable results over different accelerometer placements and different window sizes. Therefore, KNN may be considered to be a good candidate algorithm for a wide range of HAR applications.

4.7 Discussion

The field of HAR has been studied extensively. Previous literature reviews have primarily focused on aggregating common practices in HAR, with limited attention to synthesis of concrete best practice recommendations. Although it could be helpful to be aware of popular trends, practitioners should bear in mind that common practices are not necessarily well-suited. Hence, the current literature review provided a critical perspective on common practices along the HAR chain, accompanied by concrete recommendations for the design of ambulatory monitoring systems based on wearable accelerometers.

Although we agree with authors of previous reviews that no single activity recognition procedure may be universally optimal [285], we believe that several recommendations are generalisable across the HAR subdomain of ambulatory accelerometry. Consistent with previous reviews [284], the current study proposes that most practitioners aim to classify atomic activities such as walking, sitting, and lying down. Such atomic activities are relatively simple and there is a convincing body of evidence indicating that no more than two accelerometers are necessary to accurately characterise these [296–309]. So, although the generalist perspective presupposes that the number of required accelerometers for HAR depends on the complexity of activities, the design recommendation can be confined to 1-2 accelerometers by focusing on the subdomain of ambulatory monitoring.

Another recommendation which deserves additional attention concerns the degree of overlap in sliding windows. In accordance with a previous systematic review [285], sliding windows with 50% overlap were found to be most commonly used. However, until now, reviews have not addressed the adequacy of this common design choice. The results of the current study propose that the popular belief of OSWs yielding superior classification

results over NOSWs is misleading. An inherent property of existing HAR data sets is that they comprise repeated measures of the same subjects. In general, physical activities are more easily recognised by HAR systems if they concern measurements of subjects whose data have already been observed before in the training set [349]. OSWs further duplicate these repeated measures, which could lead to a false inflation of the classifier performance if practitioners do not use leave-one-subject-out CV to estimate the generalisation error [319, 326]. Overconfidence in classification performance could be harmful in clinical practice and should thus be handled with care. Therefore, practitioners should critically reflect on the implications of their design choices, rather than solely justifying them based on their common use in the research community.

The same recommendation holds true for feature selection. It was observed that many researchers resorted to domain knowledge by selecting features based on the choices reported in previous studies [283, 296–298, 301, 302, 304, 306, 307, 309, 342, 343]. However, these feature recommendations generally have limited transferability to other settings due to heterogeneity in experimental protocols [349]. For instance, recall that accelerometer outputs of physical activity recordings vary substantially depending upon the position along the human body from which they are measured [314, 315]. These location-dependent variations in accelerometer outputs subsequently affect features' values, which may alter their informativeness in classification schemes. Hence, manual feature selection through domain knowledge is not a trivial task.

The limited understanding of how features exactly contribute to HAR further complicates the task of manual feature selection. The overview of feature interpretations outlined in Table 4.2 provides at most a partial explanation of how features contribute to the induction process of HAR algorithms. The mean value of an acceleration signal, for instance, has primarily been associated with distinguishing between different sedentary behaviours through the postural information they provide [276]. Banos et al. [283], on the other hand, have demonstrated that mean accelerations from nine different sensor locations could not only distinguish between sedentary behaviours, but also between physical activities. This could be indicative of interactions between mean features, which introduce discriminative properties that are not encompassed by univariate feature interpretations as described before. Consequently, feature sets obtained through manual selection could be suboptimal due to underexploitation of feature interactions.

Feature selection algorithms such as Relief-F can accommodate for the aforementioned limitations and should be adopted in HAR more regularly to identify informative features based on their synergy, rather than solely their individual effects. However, this does not imply that domain knowledge should be fully disregarded from feature selection [350]. Studies have shown that feature selection algorithms yield increasingly random results upon increasing the cardinality of the initial candidate feature set [351]. Hence, it is recommended that researchers pre-select a candidate feature set using domain knowledge, and subsequently select the most informative subset using feature selection algorithms.

An important limitation of this review, is that all design choice recommendations were given in isolation of each other. In reality, each design choice made along the HAR chain, may affect the efficacy of another. For instance, the minimum viable window size could depend on the number of informative features included in the HAR system

[283]. Hence, the recommendations provided here should be considered to be a simplified guideline. However, when possible, evidence synthesis was based on studies which examined the influence of a design choice (e.g. the classification algorithm) under varying conditions (e.g. different sensor placements and window sizes). Therefore, some degree of generalisability can be assumed.

4.8 Conclusion

This review reports on recommendations for design choices along the HAR chain, aimed towards practitioners in the field of ambulatory accelerometry. The guidelines presented here can be used as a starting point for the development of ambulatory monitoring systems. Practioners should bear in mind that this review does not present universally optimal recommendations. It is therefore imperative that design choices are always made with careful considerations. The synthesised evidence in this review can be used to better substantiate design choices, which may lead to the development of better ambulatory monitoring systems.

CHAPTER 5

Improving Physical Activity Monitoring in Hip Fracture Rehabilitation

Abstract

Introduction: Physical activity during hip fracture rehabilitation is vital to counter long-lasting physical dysfunction amongst geriatric patients. Nevertheless, physical activity is seldom measured in clinical practice. Furthermore, existing continuous monitoring devices were developed in middle-aged adults, causing their measurements to yield unreliable results in older adults with slower gait.
Objective: This study aimed to develop a new robust human activity recognition (HAR) system to improve continuous physical activity monitoring during hip fracture rehabilitation.
Method: 24 healthy adults aged 80 years or older were included in this study. All participants' physical activities were measured in simulated free-living conditions for 75 minutes with two accelerometers: one on the lower back and one on the anterior upper thigh. Our system selected informative features from the acceleration data to classify walking, standing, sitting, lying down and postural transfers through statistical machine learning. We further tried to enhance the system's population-level impact by improving its robustness to inter-person variability. This was done by building a synthetic data set which represented the common gait characteristics of different participants to train a more generalisable classification model. Robustness to inter-person variability was measured through leave-one-subject-out cross-validation.
Results: Synthetic data showed potential to enhance classification performance and robustness to inter-person variability. The final model was able to reliably detect physical activities across most participants, with mean and standard deviations of F1-scores of 0.896 ± 0.100 for walking, 0.927 ± 0.039 for standing, 0.997 ± 0.004 for sitting, 0.937 ± 0.202 for lying down and 0.816 ± 0.120 for postural transfers. However, walking patterns associated with low acceleration amplitudes such as shuffling or slow gait remained challenging to detect.
Conclusion: The preliminary classification results in healthy adults aged 80 years or older were promising. Validation of the proposed HAR system in the hip fracture patient population remains necessary to examine the system's true utility in clinical practice.

Keywords: *hip fracture, older adults, physical activity, human activity recognition, accelerometers*

5.1 Introduction

Hip fractures are expected to cause a substantial burden on healthcare systems worldwide due to an increasing incidence in the ageing population [6]. Following surgical treatment, many older patients require rehabilitation to ameliorate physical dysfunction and to avoid bedridden state [352, 353]. Unfortunately, patients' functional recovery rates are poor, as more than half of them do not regain their prefracture mobility levels within the first postoperative year [32, 354]. This results in a loss of independence in activities of daily living (ADL) [355] and a long-lasting decline in health-related quality of life (HRQoL) [30]. Therefore, it is evident that more effective rehabilitation strategies are necessary to improve hip fractures patients' recovery.

A potential way to improve the functional recovery during rehabilitation is by implementing physical activity regimens. Various studies have demonstrated that physical activity during rehabilitation increases patients' chances of regaining their mobility and independence in ADL [39–42]. Additionally, studies have shown that physical activity counteracts the risk of secondary hip fractures [356] by preventing sarcopenia and balance deficits [357–362]. Therefore, physical activity levels are postulated to be a relevant modifiable target for intervention during rehabilitation to improve patients' functional recovery and HRQoL.

Although physical activity positively influences functional outcomes, it is seldom quantified by healthcare professionals [353]. Consequently, researchers have begun to use commercially available activity trackers, which are wearable devices which measure acceleration signals, to detect and quantify how much hip fracture patients engaged in physical activities during rehabilitation [353, 363, 364]. However, the validity of these studies' measurements could be questioned, as the employed activity tracker's reliability is known to be lower for the recognition of low-intensity activities such as (slow) walking [365]. This finding may pose concerns as geriatric hip fracture patients mostly engage in low-intensity activities [364]. Therefore, to determine whether activity trackers are sufficiently reliable to be used for rehabilitation monitoring, a validation study in the patient population was deemed necessary.

In our previous work [35], we validated whether a commercially available activity tracker with a built-in human activity recognition (HAR) algorithm could reliably measure physical activities in geriatric hip fracture patients. We found that the activity tracker's built-in HAR algorithm [366], which was calibrated to movements of middle-aged adults, severely underestimated the time spent walking in hip fracture patients aged 82 ± 6 years. Specifically, instances of walking were often misclassified as instances of standing. Since older adults exhibit distinctly slower gait patterns [317, 347, 367], algorithms developed in younger populations may fail to detect ambulation behaviours [368]. Therefore, it is necessary to develop a new monitoring system which can measure physical activities in the geriatric population more reliably.

To adequately monitor geriatric hip fracture patients during rehabilitation, it is important that the system can reliably detect all functional milestones. These are the physical activities that patients should be able to perform independently for a successful discharge, i.e. sitting, standing, lying down, walking and transfers (sit-to-stand, stand-to-sit, sit-to-lie, and lie-to-sit transfers) [369]. Research on detection of all relevant functional

milestones in the geriatric population is limited [296], as most HAR studies involving older adults overlook transfers in the development of their monitoring systems [297, 303, 317, 337, 346]. While the system developed by Allen et al. [296] can detect all milestones, its use for clinical practice may be limited since it was not developed in free-living conditions (FLC). FLC give more insights into the "diverse activity patterns due to individual habits and unpredictable real-life conditions" [285, p. 4]. To be of good practical use, HAR algorithms should be able to generalise over intra-activity variation induced by FLC [370–372]. Hence, more research into robust characterisation of the functional milestones in FLC is needed.

However, new challenges regarding the robustness of the monitoring system may arise upon using a FLC data collection protocol. Due to the lack of strict experimental instructions which prescribe how often and for how long activities should be performed, the collected data are prone to imbalances [373]. For instance, some test subjects may spend significantly more time walking than others, causing the recordings of walking to be dominated by a few subjects. Consequently, data collected in FLC may be limited in their ability to represent inter-person differences in gait. This may harm the HAR algorithm's capability to generalise well across different individuals, since inter-person variability is a major source of intra-activity variation [319, 349, 372–376]. Therefore, to enhance generalisability, novel data mining methods are needed to extract physical activity characteristics that are common across different individuals.

In this study, we aimed to develop a new monitoring system based on HAR to detect the functional milestones for hip fracture rehabilitation reliably and robustly. Firstly, to improve upon the reliability of commercially available activity trackers, the newly proposed HAR algorithm was calibrated to healthy adults aged 80 years or older, i.e. a cohort which is representative of hip fracture patients age-wise [2]. Secondly, to enhance robustness to intra-activity variation, all data were collected in simulated FLC. Thirdly, to make the HAR algorithm more robust to inter-person variability, we built an additional synthetic data set which represented the common gait characteristics of different individuals to train a more generalisable classification model.

5.2 Materials and Methods

5.2.1 Study Design

This prospective observational study was performed from May 2021 until November 2021 at the eHealth House (eHH) of the TechMed Simulation Centre at the University of Twente, Enschede, The Netherlands. The eHH is a controlled environment where FLC can be simulated. The facilities included a living room, kitchen, bedroom, and bathroom. The entire experiment was recorded with five cameras installed in the eHH.

Participants were enrolled if they were aged 80 years or older and if they were physically able to independently participate in the study. Participants were excluded if they had cognitive impairments or mobility disorders which prohibited them from participating in the study. According to the Dutch law and supported by a ruling from the appropriate ethics committee (Medical Research Ethics Committee (MREC) Arnhem - Nijmegen), the study

was exempt from the Medical Research Involving Human Subjects Act. Ethical approval for conducting the study was granted by the ethical committee Natural Sciences and Engineering Sciences of the University of Twente. Prior to participation, all participants provided written informed consent.

5.2.2 Study Procedure

The participants were invited to the eHH once for a 75-minute visit, during which they were instructed to perform several ADL tasks at their own pace in their own preferred order. These tasks were related to discharge criteria for geriatric hip fracture patients to return home and included (1) walking inside the eHH (walking back and forth to the front door, walking back and forth to the kitchen, and walking back and forth to the bedroom), (2) visiting the toilet once, (3) going in and out of bed once (sit-to-lie transfer, lying down, lie-to-sit transfer), and (4) providing meals (walking to the kitchen, cutting food and bringing it back to the living room), and (5) grabbing a drink (walking to the kitchen, pouring a drink and bringing it back to the living room). During the ADL tasks, participants' movements and postures were characterised with two movement sensing devices: the MOX activity monitor (Maastricht Instruments, The Netherlands) and the APDM activity tracker (Hankamp Rehab BV, The Netherlands).

5.2.3 Activity Trackers

The MOX is a small waterproof device comprising a single triaxial accelerometer. It was attached to the upper thigh with a special plaster, approximately 10 cm above the knee and recorded data at a sampling frequency of 25 Hz. This location was chosen based on compelling evidence that upper leg accelerations were least sensitive to inter-person differences, allowing for better generalisation in HAR [300]. The APDM comprises three small sensors: a triaxial accelerometer, a triaxial magnetometer, and a triaxial gyroscope. The APDM was worn on the lower back with a strap and recorded data at a sampling frequency of 128 Hz. This location was chosen since it yields (1) robust measurements of sedentary behaviours with low sensitivity to inter-person postural differences [303], and (2) representative characterisations of whole-body movements due to its proximity to the centre of mass of the human body [278]. Measurements from both body locations (see Figure 5.1) were deemed necessary for robust HAR since a single location was expected to be unable to provide sufficiently distinctive information for all static activities (see Appendix C.1).

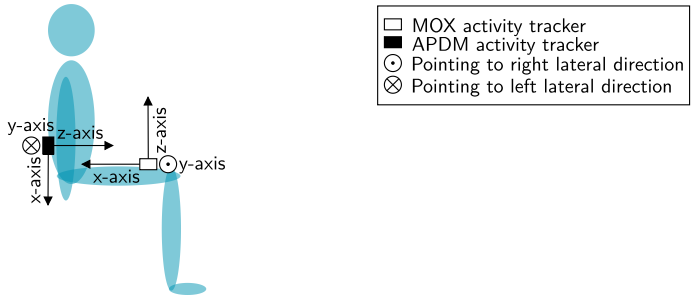

Figure 5.1: Schematic of activity trackers' placements on a test subject's upper thigh and lower back.

5.2.4 Data Annotation

To obtain gold standard labels for HAR recognition, two independent reviewers annotated the physical activities observed in the video recordings into the following categories: walking, standing, sitting, lying (supine, left and right lateral recumbent), transfers (stand-to-sit, sit-to-stand, lie-to-sit, and sit-to-lie). Discrepancies were resolved through adjudication by a third reviewer. Subsequently, through visual inspection of the acceleration data, the trackers' recordings were synchronised with the timestamps at which the first annotated sit-to-stand transfer occurred according to the videos.

5.2.5 Data Processing Pipeline

After establishing a relationship between the annotations and the activity trackers' recordings, the remaining analyses were solely carried out using the trackers' accelerometers. The magnetometer and gyroscope were excluded, since previous studies demonstrated that accelerometers were superior for HAR tasks similar to ours, and that there was no significant performance gain upon combining the sensors [377, 378]. All accelerometer data were processed using MATLAB (R2022a, MathWorks Inc., Natick, MA, USA).

To improve the HAR algorithm's robustness to inter-person variability, we built a synthetic data set which represented the common gait patterns across different individuals. In short, we proposed two classification models based on synthetic data to explore whether the common gait patterns could enhance the generalisation capabilities of HAR algorithms. We also trained a classifier based on real data to serve as a control condition model (CCM) to compare the performance with models involving synthetic data. For all three candidate models, the data processing steps were based on the activity recognition chain (ARC) [373]: preprocessing, feature extraction, feature selection, model building, and model evaluation. The individual steps are explained in more detail in the following sections.

For the first candidate model, we examined whether the synthetic data could be used to aid the classifier in learning decision boundaries for classification which were more robust to inter-person variability. The model was developed in three stages (left flow chart of Figure 5.2): feature selection, training and testing. Feature selection and training were strictly performed on the complete synthetic data set. After the model was fully trained, it was evaluated on the real data of all participants, one by one, to quantify robustness to inter-person variability. Since all steps of the model training process were performed on the synthetic data, the model was named the complete data intervention model (CDIM).

For the second candidate model, we examined whether the synthetic data could be used to aid in the identification of generalisable features which were more robust to inter-person variability. The model was developed in three stages (middle flow chart of Figure 5.2): feature selection, training and testing. During the feature selection stage, the complete synthetic data set was used to identify a feature set proposal. Subsequently, in the training phase, the proposed features were extracted from the real data to proceed with model training. The model was trained and evaluated on the real data using leave-one-subject-out cross-validation, where measurements of the left-out participant were used for testing to gain insights into robustness to inter-person variability [302].

Since we intervened in the traditional ARC by strictly performing feature selection on synthetic data, the model was named the feature intervention model (FIM).

Finally, for the CCM, we followed the traditional progression of the ARC by solely relying on real data. The CCM was developed in three stages (right flow chart of Figure 5.2): feature selection, training and testing. We used 25% of the real data to perform feature selection. Subsequently, the chosen features were used to train and evaluate the CCM using leave-one-subject-out cross-validation on the remaining 75% of the data.

Preprocessing

Firstly, random fluctuations due to noise were suppressed by smoothing all physical activity signals with a Savitzky-Golay filter [379]. A Savitzky-Golay filter was preferred over a moving average filter, as it is better at preserving the characteristic shapes and heights of peaks in signals during the smoothing process [380]. The Savitzky-Golay filter was configured with a local frame length of 0.12 seconds and a polynomial order of 2. The frame length of 0.12 was chosen as it was previously found to be suitable for smoothing gait patterns in the geriatric population [276]. The polynomial order was set to 2, as it was found to be sufficiently complex to locally capture the characteristics of movement signals recorded with accelerometers [381].

Following filtering, the data were split into non-overlapping sliding windows of 2 seconds. Non-overlapping windows were chosen since overlapping windows require more computational resources without improving a HAR system's generalisation properties according to a previous study which quantitatively compared the two window types [319]. A window size of 2 seconds was chosen as previous studies found that windows of 2 seconds generally provide low classification error rates across a wide range of different physical activities [283], and that windows of 2 seconds were particularly well-suited to capture the physical activity patterns of older adults [276]. Windows containing a mix of different activity annotations were excluded from the analysis to avoid ambiguity in classifications [311].

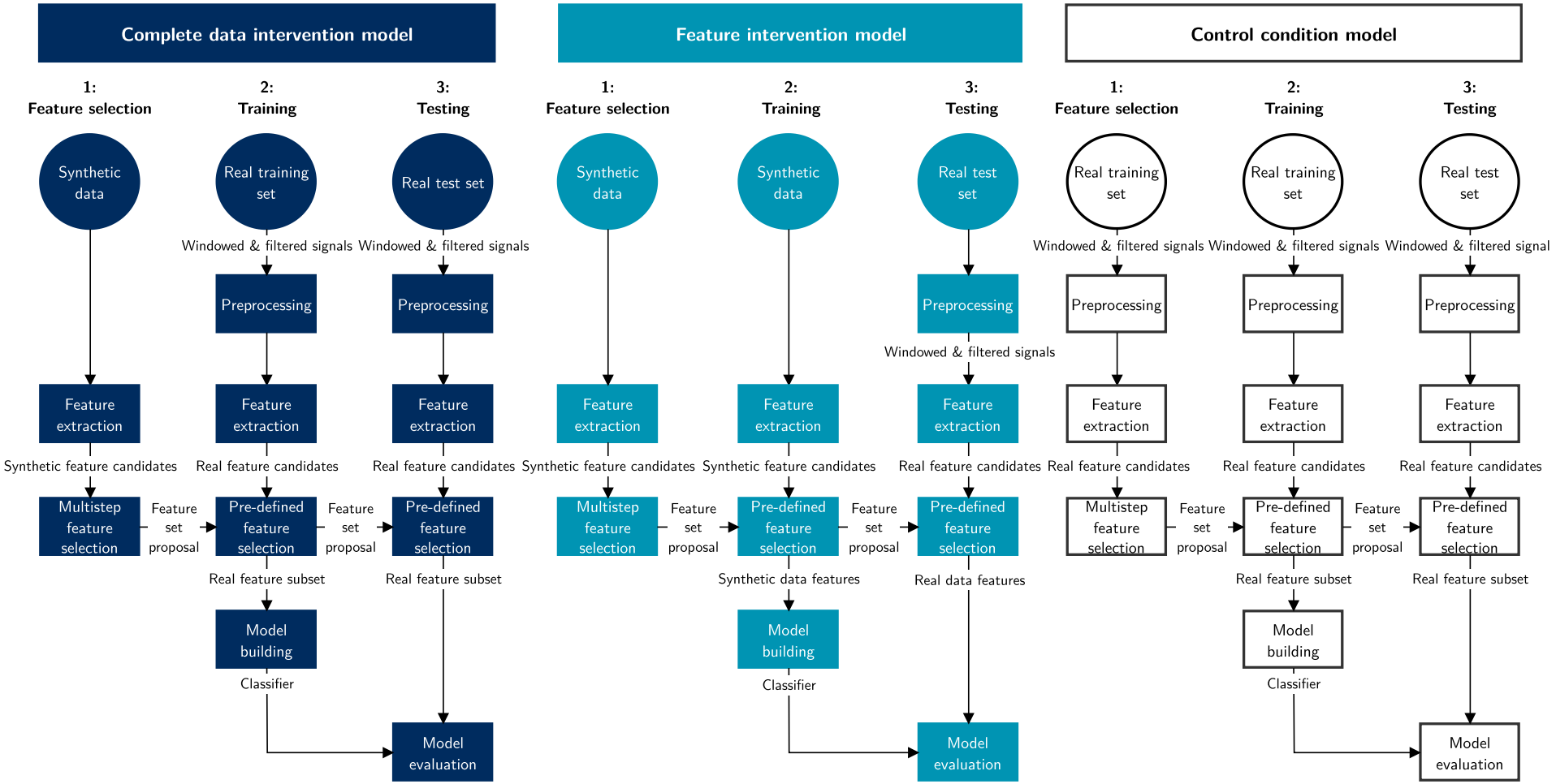

Figure 5.2: General workflow diagram of the development and evaluation of the three proposed human activity recognition algorithms. Within each of the three flow diagrams, the feature set proposals remained the same across the pre-defined feature selection steps.

Synthetic Data Generation

We generated synthetic data by finding generalised gait representations based on common characteristics across different individuals and by manually removing potential biases which could influence feature selection and classification algorithms. Firstly, to remove biases, the raw acceleration data of all activities were visually examined for anomalies. For walking, unusually large baseline drifts along the z-axis of the APDM were observed amongst a few participants (see Appendix C.2 for an example). Since no kinematically plausible cause could be determined, the drifts were deemed anomalous and we removed them to streamline the signal trends with those observed in the majority of participants. This was done by applying a noncausal fourth-order high-pass filter with a cut-off frequency of 0.5 Hz. Secondly, to generate gait representations based on common characteristics across different individuals, the dynamic time warping barycentre averaging (DBA) [382] algorithm was used. We generated synthetic counterparts for each of the following activities: walking, standing, sitting, lying supine, lying down in left lateral recumbent position, lying down in right lateral recumbent position, sit-to-stand transfer, stand-to-sit transfer, sit-to-lie transfer, and lie-to-sit transfer. We kept the number of available 2-second windows for each of these activities identical across both the real and synthetic data, to fairly test the potential merit of using synthetic data in identical conditions of class imbalance.

DBA has been widely used in signal processing and bioinformatics to extract common characteristics from a set of time series which may have variations in timing or pace [382–389]. It is based on the dynamic time warping (DTW) algorithm [390], which aligns multiple time series by nonlinearly stretching or compressing the time axes to minimise the difference in the values of the time series at each point in time. A key boundary condition imposed during the alignment process is that the first and last points of the compared time series must always be matched. An example of what alignment under DTW looks like is shown in Figure 5.3. Following DTW, DBA then calculates the average of the aligned time series to obtain a representative time series which captures the common features, regardless of variations in timing or pace.

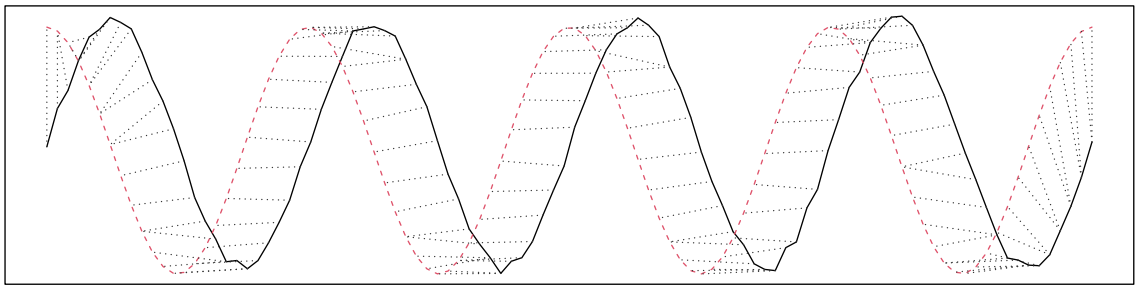

Figure 5.3: Example of two sequences which have been matched and aligned through dynamic time warping. The black dashed lines depict the matched points between the two sequences.

To apply DBA to our study, we defined a custom protocol to generate synthetic activity windows of 2 seconds (see Figure 5.4). The protocol comprised four main steps. Firstly, for a given activity, we determined whether sufficient recordings (>100 windows) were available to construct a generalised and diverse synthetic data set. If this was the case,

one activity window was sampled from each participant as input for DBA. When fewer windows were available, windows were subsampled across 20% of the participants to ensure sufficient variability in the synthetic data. This rule of thumb was established since we anticipated approximately 25 participants to participate in the study. If every iteration of DBA then used 25 windows for synthetic data generation and fewer than 100 windows were available in the sampling space, we anticipated that the synthetic windows would look too similar. By subsampling across five participants (20%), the lack of diversity in the synthetic data due to the limited number of available windows could be partially countered, since there are 53,130 unique combinations of selecting five participants from a group of 25.

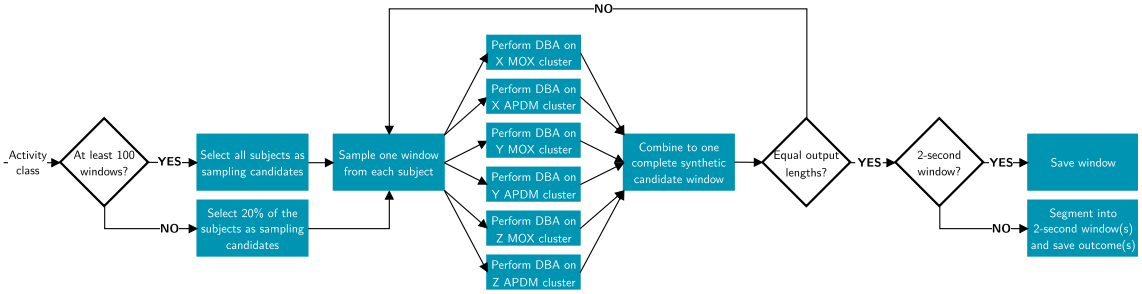

Figure 5.4: General protocol for the generation of synthetic data using dynamic time warping barycentre averaging (DBA).

Secondly, after establishing an adequate selection of participants, we proceeded with sampling activity windows which served as inputs for DBA. For each selected participant, one window was randomly sampled. One window was defined as a set of simultaneous measurements from the MOX and APDM along the x-, y-, and z-axes for a given time interval. The interval width, i.e. the window size, was customised for different activities to mitigate artefacts at the start and end of synthetic signals due to violations of the boundary conditions of DTW. For walking, we exploited its periodicity by providing longer inputs than necessary (4 seconds). Potential violations of the boundary conditions could then be ignored by solely extracting the 2-second segment of the 4-second window which was free of artefacts. For transfers, which were not periodic and varied in duration, we did not choose a fixed window size. Instead, transfers were processed from start to end, irrespective of length, to ensure that the boundary conditions were satisfied. For static activities, we directly processed 2-second windows without concerns regarding boundary artefacts. The concerns were neglected as static activities did not have distinctive boundaries due their minimal changes in acceleration content from start to end.

Thirdly, after sampling windows of adequate length for the respective activity, we attempted to generate a temporally consistent synthetic window using DBA. To generate a single window, the DBA algorithm was ran six times, once for each of the x-, y-, and z-axes recordings of the MOX and APDM separately. A synthetic window was said to be temporally consistent, if all six signals were of equal duration in time. This verification was only necessary for transfers, since the input lengths were allowed to vary here. Specifically, DTW solved the optimal alignment problem by stretching and compressing time axes, such that the length of the synthetic signal produced by DBA always varied between the

length of the shortest and longest input sequence. Depending upon whether the optimal alignment was found by primarily stretching or primarily compressing the time axes, some of the synthetic accelerations could be shorter than others. In case that the six synthetic signals of a single window were temporally inconsistent, the window sampling step was repeated to inform a new iteration of DBA.

Finally, once a temporally consistent synthetic window had been generated, we proceeded with the final post-processing stage to complete the generation of a 2-second window: truncation and segmentation. The choice between truncation, segmentation and no post-processing was dependent upon the window sizes chosen in the second step of our custom DBA protocol. For walking, the first and last seconds of the 4-second window were truncated to obtain a 2-second window that was free of DTW boundary artefacts. For transfers with durations longer than 2 seconds, the windows were segmented into non-overlapping 2-second windows. For static activities, no post-processing was needed since all windows were already 2 seconds long.

Feature Extraction

For the feature extraction step, the choice was made to rely on handcrafted features. These are defined as discriminative attributes that researchers manually compute from an activity window based domain knowledge. Although automatic feature extraction methods based on deep learning are postulated to yield the most promising results with regards to generalisability [288–294], they were deemed infeasible as large volumes of data are a prerequisite for meaningful results [370]. Since the functional milestones we aimed to characterise included transfers, which are well-known to have low occurrence rates in HAR data [391–393], methods based on deep learning were considered unsuitable.

The candidate set of handcrafted features was identified based on literature. Although time and frequency domain features are both commonly used in HAR [341], we limited ourselves to time domain features for multiple reasons. Firstly, there is no compelling evidence that frequency domain features yield superior discriminative properties in exchange for higher computational costs [296, 300, 394]. Secondly, frequency domain features primarily lend themselves useful for the characterisation of quasi-periodic movements with distinct frequency content [296]. Amongst the physical activities of interest, most were either static or aperiodic. Hence, frequency domain features were anticipated to be of minimal utility and thus removed preemptively.

Additionally, a subselection was made amongst the time domain features to reduce multicollinearity. For instance, it was anticipated that the mean and median values of acceleration signals would provide nearly identical information about physical activities, causing inclusion of both to be redundant. Similar outcomes were expected for the standard deviation, mean absolute deviation, and variance. This resulted in a total of 62 candidate features as shown in Table 5.1. By preemptively eliminating redundant features based on domain knowledge, as recommended by multiple researchers [350, 395], subsequent feature set optimisations through feature selection (FS) algorithms can be handled reliably [351].

Table 5.1: Overview of the time domain features considered in the initial candidate set. Each feature was computed for both the MOX and APDM signals across a window of size N. Except for the axial correlations and signal vector magnitude, all features were computed along each individual acceleration axis $a \in \{x, y, z\}$. This resulted in a total of 62 features.

Feature and interpretation	Formula	References
The **mean** provides postural information. It can be used to distinguish between different sedentary behaviours.	$\mu_a = \frac{1}{N}\sum_{i=1}^{N} a_i$	[276, 283, 296, 298, 300–309, 313, 328, 329, 331]
The **standard deviation** provides information about the intensity of a physical activity. It can be used to distinguish between sedentary behaviours and physical activities.	$\sigma_a = \sqrt{\frac{\sum_{i=1}^{N}(a_i - \mu_a)^2}{N-1}}$	[298, 300, 301, 306, 307, 309, 313, 328, 329, 331]
The **root mean square** is identical to the standard deviation for signals with a mean of zero.	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i}$	[297, 298, 300, 306, 331]
Axial correlations can be used to distinguish between activities which involve unidirectional and multidirectional translations. Walking and running, for instance, primarily involve unidirectional translations (anteroposterior), whereas stair climbing involves prominent multidirectional translations (anteroposterior and vertical).	$\frac{cov(x,y)}{\sigma_x\sigma_y}$, $\frac{cov(x,z)}{\sigma_x\sigma_z}$, $\frac{cov(y,z)}{\sigma_y\sigma_z}$	[297, 300, 304–307, 312, 329, 331]
The **minimum** describes the lowest value measured along a single sensing axis of an accelerometer.	$min(a_1, a_2, ..., a_N)$	[297, 298, 300, 302, 303]
The **maximum** describes the highest value measured along a single sensing axis of an accelerometer.	$max(a_1, a_2, ..., a_N)$	[297, 298, 300, 302, 303]
The **interquartile range** is the difference between the 25% and 75% quantiles of the acceleration values. It can be used to distinguish between sedentary behaviours and physical activities.	$Q_3 - Q_1$, where Q_3 is the 75% quantile and Q_1 is the 25% quantile.	[300, 331]
The **skewness** provides information about the shape of the probability distribution of movement accelerations. It can take on negative and positive values. If the values of the acceleration signals are symmetrically distributed, the skewness is equal to zero. If high values are more likely to occur than low values, the skewness is positive-valued.	$\frac{\sum_{i=1}^{N}(a_i - \mu_a)^3}{(N-1)\sigma_a^3}$	[298, 300, 307, 328, 329, 331]
The **kurtosis** provides information about the shape of the probability distribution of movement accelerations. It describes how likely extreme values occur in a movement signal.	$\frac{\sum_{i=1}^{N}(a_i - \mu_a)^4}{(N-1)\sigma_a^4}$	[298, 300, 307, 328, 329, 331]
The **mean-crossing rate** provides information on how fast accelerations change directions during movements. It can be used to distinguish between dynamic physical activities. It is computed using the $sgn(b)$ function, where $sgn(b)$ = -1 for $b < 0$ and +1 for $b > 0$.	$\frac{\sum_{i=2}^{N} \lvert sgn(a_i - \mu_a) - sgn(a_{i-1} - \mu_a)\rvert}{2}$	[283, 312, 329, 331]
The **signal vector magnitude** (SVM) can be used to distinguish sedentary behaviours from physical activities. All accelerations were filtered with a noncausal fourth-order high-pass filter with a cut-off frequency of 0.5 Hz prior to computing the SVM.	$\frac{1}{N}\sum_{i=1}^{N}\sqrt{x_i^2 + y_i^2 + z_i^2}$	[300, 341]

Multistep Feature Selection

To improve the generalisability of the HAR classification model, we proposed a novel FS pipeline to select a minimal set of features from the initial candidate set. The overall procedure, which was applied to both real and synthetic data, is summarised in Figure 5.5. In short, the pipeline addressed two generalisability challenges in FS. The first challenge was the absence of a single universally optimal FS algorithm, as the effectiveness of an individual FS algorithm's heuristics is highly application-dependent [396, 397]. We addressed this by designing a heterogeneous feature selection ensemble (HFSE) [396]. A HFSE uses multiple FS algorithms to identify relevant features based on a diverse and complementary range of selection heuristics, causing the final feature set to be more generalisable [396–400]. The second challenge concerned the robustness of the HFSE to small changes in the data [395]. This was addressed by imposing FS stability as a secondary selection criterion. As proposed in [401], we repeated FS 10 times on activity-stratified subsamples of the data with approximately 78% overlap. The 10 outcomes were combined through vote counting: features were included in the final set if they were selected by the HFSE at least 5/10 times. This threshold was chosen as it is commonly used to eliminate features that were selected by chance [402–404].

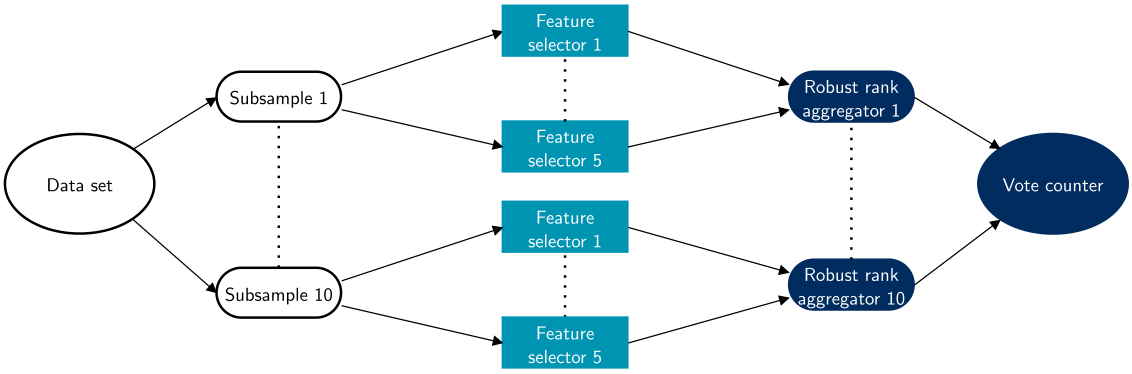

Figure 5.5: Overview of the multistep feature selection pipeline. The pipeline was applied to the real data to build the control condition model, and it was applied to the synthetic data to build the complete data intervention model and the feature intervention model.

Designing the HFSE comprised two steps: selecting the FS algorithms for the ensemble, and choosing the aggregation method to combine the ranking lists. For the first design step, five FS algorithms were selected: Relief-F [345], maximum relevance minimum redundancy (MRMR) [405], interaction-curvature tests embedded into a decision tree [406], out-of-bag (OOB) feature importance by permutation embedded into a random forest [407], and regularisation embedded into a linear discriminant [408]. The selection of these FS algorithms was based on the two primary criteria for well-performing HFSE: diversity in selection heuristics and stability of the individual FS algorithms described above [396, 399].

Firstly, diversity was ensured by composing an ensemble such that the limitation of one FS algorithm could be overcome by the strength of another (see Table 5.2). For instance, Relief-F is known to be one of the few FS algorithms which is capable of capturing interactions between features. However, it is unable to identify redundancy in feature sets.

MRMR, on the other hand, does consider feature redundancy but it neglects interactions between features.

Table 5.2: Overview of the individual feature selection algorithms used in the heterogeneous feature selection ensemble, accompanied by their properties.

Algorithm	Type	Relationship	Feature redundancy	Classifier type
Relief-F	Filter	Multivariate	Not accounted for	-
MRMR	Filter	Univariate	Accounted for	-
ICT[a]	Embedded	Multivariate	-	Nonlinear & greedy
OOB importance[b]	Embedded	Univariate	-	Nonlinear & non-greedy
Regularisation[c]	Embedded	Univariate	-	Linear

MRMR maximum relevance minimum redundancy, *ICT* interaction-curvature tests, *OOB* out-of-bag
[a] Embedded into decision tree classifier
[b] Embedded into random forest classifier
[c] Embedded into linear discriminant classifier

Secondly, we empirically verified that the five FS algorithms were sufficiently stable to reliably contribute to the HFSE. Here, stability referred to the consistency of FS algorithms' results under variations of the data. To assess this, each FS algorithm was evaluated on 10 activity-stratified subsamples of the synthetic data with approximately 78% overlap to produce 10 feature ranking lists. Following common practices in literature [396, 401, 409], the Tanimoto similarity T (5.1) was chosen to measure the FS stability across the 10 lists, where $\{s, s'\}$ denotes a pair of feature subsets obtained from the same FS algorithm applied to two different subsamples, $|.|$ denotes the feature set cardinality, and $s \cap s'$ denotes the intersection between the two subsets. The Tanimoto similarity quantifies the overlap of features between subsets s and s' regardless of their rankings, with $T = 0$ indicating no overlap and $T = 1$ indicating identical subsets. Since most of our FS algorithms solely ranked features by importance without selecting a subset, we defined the subsets s and s' as the top 10 features of a ranking list to compute T as proposed in [401]. T was computed for all 45 unique pairs of $\{s, s'\}$ that could be constructed from the 10 feature ranking lists. The overall stability for a single FS algorithm was then determined by averaging over all 45 estimates.

$$T(s, s') = 1 - \frac{|s| + |s'| - 2|s \cap s'|}{|s| + |s'| - |s \cap s'|} \tag{5.1}$$

For the second design step, the robust rank aggregation (RRA) algorithm [410] was chosen to combine the ranking lists produced by the individual FS algorithms. This model-based approach uses order statistics to forge a consensus set by selecting features which ranked significantly better than expected by chance ($p < 0.05$). Aggregation techniques based on order statistics were preferred over simpler methods such as rank averaging, as these simpler methods tend to produce erratic results [397], causing their performance to be inferior [410]. Amongst the available methods based on order statistics, i.e. RRA and Stuart's method [411, 412], RRA was preferred based on its substantially lower false discovery rate (FDR) [410]. A low FDR was prioritised to prevent overfitting by limiting the feature set cardinality [413].

In short, the RRA algorithm functions as follows. Each feature was associated with five ranks: one estimated by each FS algorithm in the HFSE. Each rank varied between 1 and m, where m denotes the total number of examined features. For each feature, RRA first normalised each rank over m, such that $r_i \in (0, 1]$, for $i = 1, 2, 3, 4, 5$. Let the collection of normalised ranks for a single feature be denoted by $\boldsymbol{r} = (r_1, r_2, r_3, r_4, r_5)$ and let $r_{(1)}, r_{(2)}, r_{(3)}, r_{(4)}, r_{(5)}$ be a reordering of $\boldsymbol{r}$ which satisfies $r_{(1)} \leq r_{(2)} \leq r_{(3)} \leq r_{(4)} \leq r_{(5)}$. The RRA algorithm examined how probable it was to obtain $\hat{r}_{(k)} \leq r_{(k)}$ if all $\hat{r}_i \in \hat{\boldsymbol{r}}$ were sampled from the uniform distribution $\mathcal{U}(0, 1)$. Under this null model, the probability that the order statistic $\hat{r}_{(k)}$ was smaller than or equal to $r_{(k)}$, was estimated by the binomial probability $\beta_{(k)}$ (5.2). The p-value indicating whether a feature ranked significantly better than expected by chance was estimated as the minimum $\beta_{(k)}$ associated with $\boldsymbol{r}$ with a post-hoc Bonferroni correction (5.3).

$$\beta_{(k)} := \sum_{\ell=k}^{5} \binom{5}{\ell} r_{(k)}^{\ell} \left(1 - r_{(k)}\right)^{(5-\ell)} \tag{5.2}$$

$$p(\boldsymbol{r}) = \min\left(1, 5 \cdot \underset{k \in \{1,2,3,4,5\}}{\arg\min}\ \beta_{(k)}(\boldsymbol{r})\right) \tag{5.3}$$

Model Building

For the CCM, CDIM, and FIM, the classification algorithm and hyperparameters were kept fixed to allow for fair comparisons. K-nearest neighbours (KNN) was selected as the classification algorithm, since it consistently performed well across various HAR studies [299–301, 313], generally required fewer features and smaller window sizes to attain a high classification performance [283], and remained superior across classifier comparisons under varying sensor placements [300].

We abstained from hyperparameter tuning, since it was deemed infeasible to partition the data into meaningful tuning sets due to severe class imbalance and low availability of data from the minority class. Since the optimal hyperparameters for KNN have been studied extensively, the hyperparameters were based on evidence from literature. Firstly, studies have repeatedly demonstrated that $k \in [3, 10]$ works well for various HAR tasks, and that variations in this range minimally affect KNN's performance [283, 299, 300, 311, 330, 378]. From the empirical range reported in literature, we conservatively chose $k = 5$ based on several theoretical considerations. Firstly, the lower bound reported in literature was not chosen since KNN is known to be more prone to overfitting for small values of k [414]. Secondly, it is still desired to choose k to be small, as the good performance of KNN lies in its ability to capture small local differences [300]. Thirdly, upon considering values of k that were slightly larger than 3, $k = 5$ was preferred over $k = 4$, as an odd number of neighbours would avoid a tie in majority voting [378]. For the distance metric, the Euclidean distance was chosen as it was found to be superior according to one of the most extensive inquiries into hyperparameter tuning for KNN in HAR to date [300]. Finally, since KNN is sensitive to feature ranges due to its distance-based similarity heuristics, we configured KNN with Z-standardisation of the features to ensure that all features could contribute fairly to classifications regardless of their scales [299].

Classification and Evaluation

The CCM, CDIM, and FIM were trained to recognise the following five activity classes: walking, standing, sitting, lying down and transfers. Due to limited data for each of the individual four transfers (sit-to-stand, stand-to-sit, sit-to-lie and lie-to-sit), all four were merged into a single activity class. For performance comparison, the F1-scores (5.4) were computed for each model. The best performing model was chosen by comparing the mean and standard deviations of the models' F1-scores across evaluations on individual participants.

$$\begin{aligned} \text{F1-score} &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \\ \text{precision} &= \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \\ \text{recall} &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \end{aligned} \tag{5.4}$$

5.3 Results

5.3.1 Data Set

Table 5.3: Overview of how many 2-second windows were available for each physical activity.

Physical activity	Available windows		N	Windows per participant	
	Abs. freq.	Rel. freq.		Median	Range across N
Walking	2,244	8.0%	24	90.5	(58 - 139)
Standing	3,258	11.6%	24	133	(36 - 234)
Sitting	20,677	73.5%	24	967	(322 - 1,487)
Lying down (supine)	929	3.3%	21	19	(5 - 171)
Lying down (left lateral recumbent)	300	1.1%	5	24	(15 - 206)
Lying down (right lateral recumbent)	311	1.1%	2	155.5	(6 - 305)
Sit-to-stand transfer	178	0.6%	24	8	(2 - 10)
Stand-to-sit transfer	163	0.6%	24	7	(3 - 11)
Sit-to-lie transfer	35	0.1%	23*	1	(1 - 3)
Lie-to-sit transfer	41	0.1%	24	1.5	(1 - 4)

N number of participants, *abs. freq.* absolute frequency, *rel. freq.* relative frequency
* The sit-to-lie transfer of one participant was missing as the activity could not be verified with the video recordings due to a faulty camera orientation

In total, 24 participants were included in this study. This sample comprised 11 male and 13 female participants. The median age and interquantile range of this cohort were 82 and (81-85) years respectively. A brief summary of the data set is shown in Table 5.3. It is evident that data collection in a simulated FLC contributed to data imbalance in three ways. Firstly, class imbalance was present, where transfers were the minority class (1.4%) and sitting was the majority class (73.5%). Secondly, there was imbalance in the number of participants contributing to physical activity measurements. For instance, for left and

right recumbent positions, measurements of only five and two participants were available respectively. Finally, there was imbalance in measurements per participants. For lying down in a right lateral recumbent position, for instance, one participant only contributed six windows to the data set, whereas another participant contributed 305 windows.

5.3.2 Feature Selection

All five individual FS algorithms were found to be sufficiently reliable to contribute to the HFSE, since their Tanimoto similarity scores varied between 0.648-1.0 which indicated good to excellent stability [396]. The results of the HFSE applied to 10 subsamples of the real data and 10 subsamples of the synthetic data are shown in Figure 5.6. Across the 10 subsamples of the real data, three features were stably selected for the CCM. Features from both the upper thigh (UT) and lower back (LB) accelerations were selected, which were mean accelerations (Mean-X UT, Mean-Z UT), and minimum accelerations (Min-X LB).

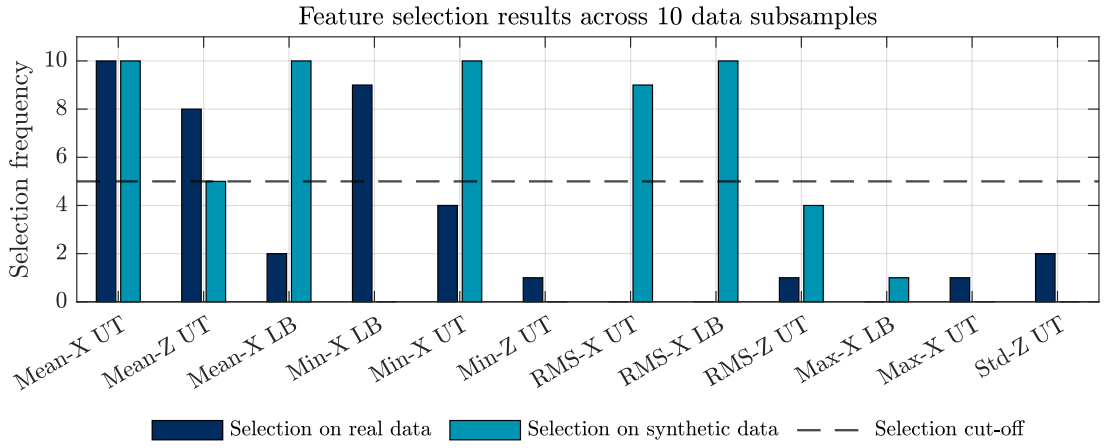

Figure 5.6: Overview of feature selection results of the heterogeneous feature selection ensemble across the 10 subsamples of the real and synthetic data.

Across the 10 subsamples of the synthetic data, six features were stably selected for the CDIM and FIM. The selected features comprised mean accelerations (Mean-X UT, Mean-Z UT, Mean-X LB), minimum accelerations (Min-X UT), and root mean squares of accelerations (RMS-X UT, RMS-X LB). It was noteworthy that the two mean features were the only ones to overlap with the feature set selected based on the real data.

The class conditional distributions of the features selected for the CCM, CDIM, and FIM are shown in Figure 5.7. Overall, the features showcase good (pairwise) separability between the different physical activities in both the real and synthetic data. For instance, the mean accelerations generally showcased good pairwise separability between the different static postures. Furthermore, the feature values of Min-X UT and RMS-X UT appeared to be rather distinctive between standing and walking.

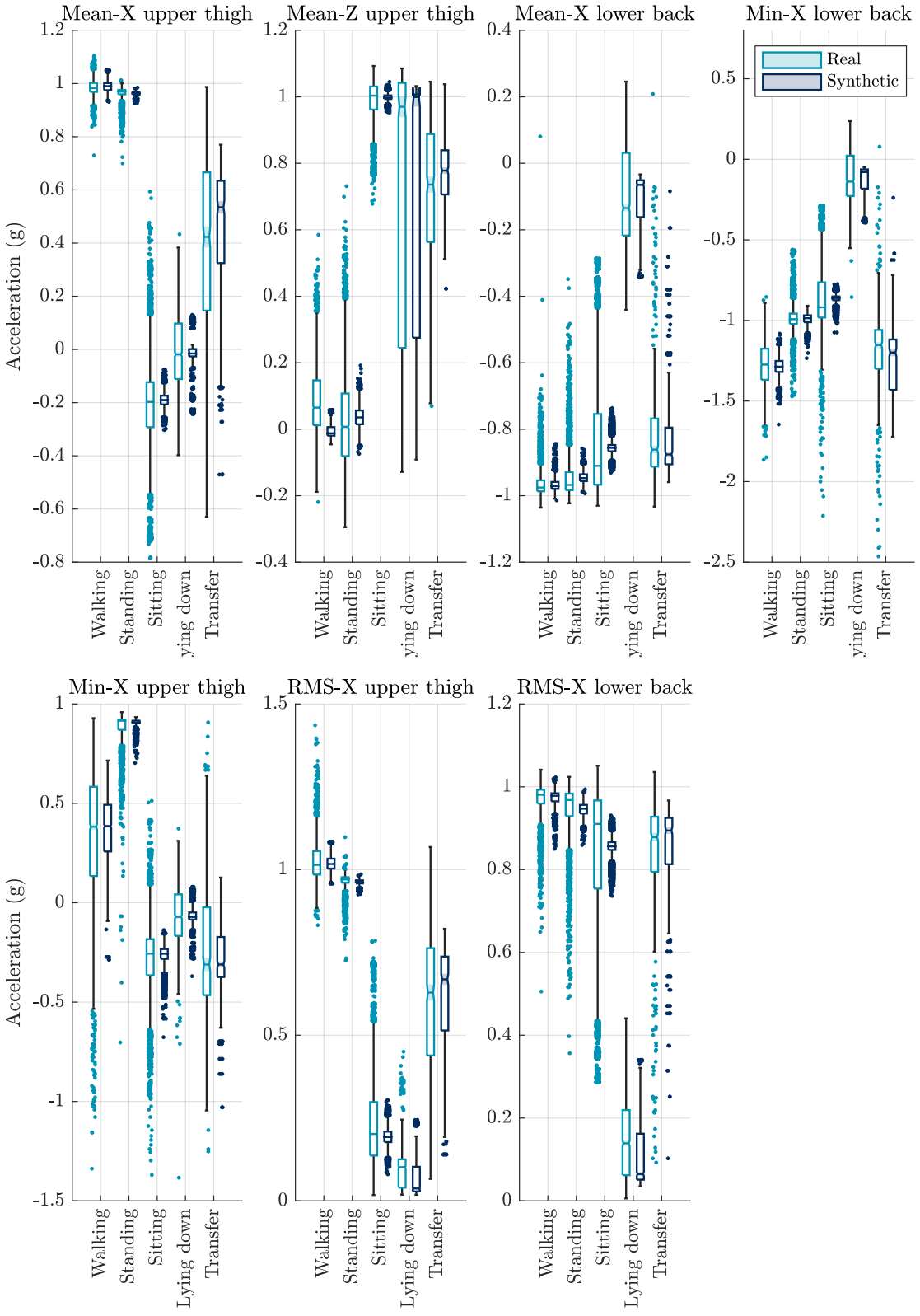

Figure 5.7: Comparison of feature distributions across the real and synthetic physical activity signals.

Generally, the degree of pairwise separability of feature values between different physical appeared to be greater across the synthetic data. The reason for this was that features

extracted from the synthetic data showcased less variability than features extracted from the real data, leaving less room for the feature ranges to overlap. This entailed that the synthetic data exhibited a lower degree of intra-activity variation. Nevertheless, there was no obvious difference between the real and synthetic data according to the box plots and case-by-case comparisons of real and synthetic data (see Figure 5.8). This indicated that the synthetic data could present the common patterns of physical activities.

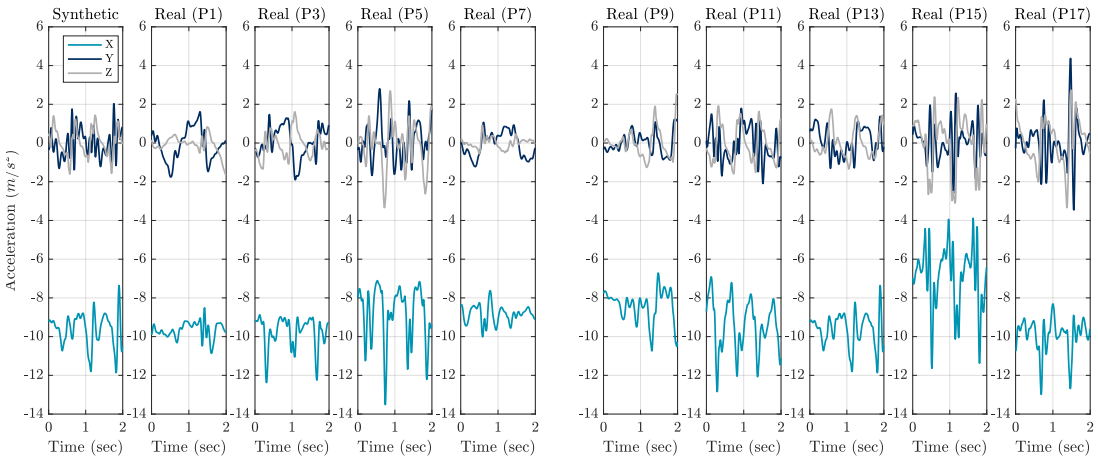

Figure 5.8: Comparison of synthetic and real physical activity signals for walking. Data from all 24 participants were used to generate the synthetic signals. Examples of real physical activity signals for 9/24 participants are shown here to give a general impression.

5.3.3 Model Evaluation

Control Condition Model

The classification performance of the CCM estimated through leave-one-subject-out cross-validation is shown in Table 5.4. Overall, HAR was performed with a mean precision of 0.932±0.045, a mean recall of 0.885±0.064, and a mean F1-score of 0.896±0.075 across all 24 participants. Upon examining the recognition rates of the individual activities, inter-person variability in classification performance appeared to be large for recognition of walking (F1-score range: 0.107-1.0) lying down (F1-score range: 0.154-1.0) and transfers (F1-score range: 0.176-0.929).

Table 5.4: Performance of control condition model evaluated with leave-one-subject-out cross-validation.

Physical activity	**Precision**		**Recall**		**F1-score**	
	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range
Walk	0.894±0.072	0.744-1.0	0.861±0.197	0.058-1.0	0.861±0.172	0.107-1.0
Stand	0.914±0.070	0.708-1.0	0.912±0.086	0.654-1.0	0.911±0.066	0.680-1.0
Sit	0.994±0.006	0.977-1.0	0.999±0.002	0.993-1.0	0.997±0.003	0.987-1.0
Lie down	0.986±0.037	0.833-1.0	0.956±0.187	0.083-1.0	0.955±0.172	0.154-1.0
Transfer	0.871±0.187	0.100-1.0	0.698±0.130	0.438-0.867	0.755±0.159	0.176-0.929

Complete Data Intervention Model

The aggregated classification performance of the CDIM evaluated per participant is shown in Table 5.5. Overall, HAR was performed with a mean precision of 0.828±0.101, a mean recall of 0.881±0.070, and a mean F1-score of 0.812±0.113 across all 24 participants. Upon examining the recognition rates of the individual activities, inter-person variability in classification performance appeared to be large for recognition of sitting (F1-score range: 0.339-0.999), lying down (F1-score range: 0-1.0), and transfers (F1-score range: 0.034-0.929).

Table 5.5: Performance of complete data intervention model.

Physical activity	**Precision**		**Recall**		**F1-score**	
	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range
Walk	0.918±0.040	0.845-1.0	0.897±0.092	0.596-0.984	0.826±0.118	0.600-1.0
Stand	0.925±0.057	0.727-0.992	0.937±0.049	0.815-1.0	0.930±0.040	0.800-0972
Sit	0.998±0.002	0.995-1.0	0.837±0.252	0.197-1.0	0,884±0.201	0.329-0.999
Lie down	0.892±0.279	0-1.0	0.910±0.282	0-1.0	0.900±0.279	0-1.0
Transfer	0.407±0.384	0.017-1.0	0.826±0.118	0.600-1.0	0.444±0.342	0.034-0.929

Feature Intervention Model

The classification performance of the FIM estimated through leave-one-subject-out cross-validation is shown in Table 5.6. Overall, HAR was performed with a mean precision of 0.936±0.055, a mean recall of 0.905±0.066, and a mean F1-score of 0.915±0.064 across all 24 participants. Upon examining the recognition rates of the individual activities, inter-person variability in classification performance appeared to be large for recognition of lying down (F1-score range: 0-1.0), and moderately large for walking (F1-score range: 0.448-0.966) and transfers (F1-score range: 0.500-1.0).

Table 5.6: Performance of the feature intervention model evaluated with leave-one-subject-out cross-validation.

Physical activity	**Precision**		**Recall**		**F1-score**	
	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range
Walk	0.922±0.037	0.855-1.0	0,889±0.138	0.288-1.0	0.896±0.100	0.448-0.966
Stand	0.924±0.051	0.768-1.0	0.932±0.051	0.769-1.0	0,927±0.039	0.833-0.975
Sit	0.994±0.008	0.961-1.0	0.999±0.001	0.996-1.0	0.997±0.004	0.980-1.000
Lie down	0.929±0.205	0-1.0	0.948±0.204	0-1.0	0.937±0.202	0-1.0
Transfer	0.909±0.136	0.417-1.0	0.758±0.151	0.500-1.0	0.816±0.120	0.500-1.0

Performance Comparison

On average, the FIM achieved the highest overall F1-score with the lowest inter-person variability (0.915±0.064), followed by the CCM (0.896±0.075) and CDIM (0.828±0.101). The CDIM was deemed unsuitable for clinical monitoring, based on its poor transfer

recognition rate with high sensitivity to inter-person variability (F1-score: 0.444±0.342). Amongst the remaining two candidate models, the FIM was preferred. The superiority of the FIM compared to the CCM was most noticeable in the classification performance for transfers on an individual participant level. The FIM improved the F1-score of transfer recognition for 18/24 participants, while retaining identical performance for 1/24 participants and only decreasing the performance for 5/24 participants (see Appendix C.3 for more details). Overall, the F1-score range across all participants increased from 0.176-0.929 to 0.500-1.0.

Inter-Person Variability Biases in the Best Classification Model

The FIM still produced relatively poor F1-scores (< 0.5) for lying down and walking in two participants. Specifically, all 12 cases of lying down were misclassified as sitting for participant 18, and 33/48 cases of walking were misclassified as standing for participant 1. To illustrate which forms of inter-person variability in gait caused generalisation issues, we compared the gait patterns of participants 18 and 1 with those of the two best classified participants for lying down and walking respectively.

First, the misclassifications for participant 18 are discussed. As previously observed in Figure 5.7, the mean acceleration feature (Mean-X LB) carried distinctive information to distinguish between the postural differences of sitting and lying down. Figure 5.9 showed that most participants, including the best classified participant, exhibited Mean-X LB values between -0.1-0.1g. This acceleration range entailed that the vertical axis of the upper body was oriented (nearly) parallel to the bed's surface while lying down. For participant 18, the Mean-X LB values were more deviant with fluctuations around -0.4g, which were in the vicinity of the Mean-X LB range observed during sitting.

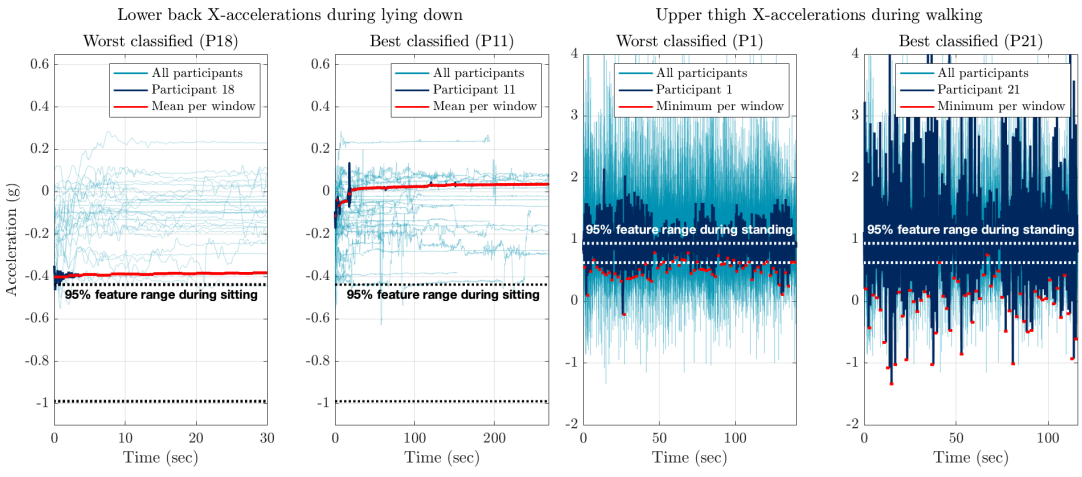

Figure 5.9: Comparison of worst and best classified individuals across physical activities with the most substantial inter-person variability. Amongst the least accurate classifications, lying down was mistaken for sitting, and walking was mistaken for standing. The 95% feature ranges describe the ranges within which 95% of the class conditional feature values lie, i.e. the 2.5% and 97.5% quantiles.

Next, the misclassifications for participant 1 are discussed. Based on Figure 5.9, it can be observed that the walking accelerations of participant 1 exhibit considerably smaller

amplitudes than those of most participants, and the difference is especially noticeable compared to the best classified participant. As previously observed in Figure 5.7, the minimum acceleration feature (Min-X UT) was found to separate most cases of walking and standing from each other. However, as illustrated in Figure 5.9, the low amplitudes of participant 1 caused the Min-X UT values during walking to fall within the Min-X UT value range observed during standing.

5.4 Discussion

This study aimed to develop a robust HAR algorithm, which could be used to continuously monitor physical activity in geriatric hip fracture patient rehabilitation. To improve the system's robustness to intra-activity variation, we collected all data in simulated FLC. To improve the system's robustness to inter-person variability in gait patterns, we created a synthetic data set which emphasised common gait patterns across different individuals. We found that the synthetic data supported our novel multistep feature selection pipeline in identifying features which generalised better across different individuals. By using the proposed feature set selected from the synthetic data to train a classifier on the real data, i.e. the FIM, the classification performance and robustness to inter-person variability improved slightly compared to the CCM. The aforementioned improvements were most noticeable in the classifications of transfers.

The differences between the real and synthetic data are worth examining to infer how they complemented each other in the development of the FIM. We previously observed that the synthetic data exhibited a lower intra-activity variation than the real data. It is postulated that this was due to a reduced inter-person variability, since DBA produced generalised gait patterns based on the common characteristics across different individuals. We believe that the elimination of individual gait deviations could be seen as a form of noise suppression, which aided the HFSE in identifying features which characterised the fundamental patterns of physical activities. This is consistent with Bai et al. [317], who found that reduced variability in HAR data could improve the stability of pattern recognition in machine learning. However, our findings show that the variability reduction in the synthetic data only improved pattern recognition during feature selection, since the CDIM performed relatively poorly. Hence, exposure to real-life intra-activity variation during training remains necessary to ensure more generalised decision boundaries for classification.

With the FIM, we were able to classify most physical activities correctly using solely six features selected with our proposed multistep feature selection method. These were: Mean-X UT, Mean-Z UT, Mean-X LB, Min-X UT, RMS-X UT, and RMS-X LB. Since the FIM only relies on a few features, it has the advantage that the predictors for the activity classifications can be pinpointed effectively. In Figure 5.7, we provided an overview of the feature value distributions of these six features across each physical activity, to illustrate how they could make distinctions between pairs of physical activities from a univariate point of view. From here, several lessons could be learnt:

Firstly, we observed that the mean accelerations primarily supported the discrimination between standing, sitting, and lying down. These findings are in line with previous HAR

studies which characterised body postures using the mean. For example, Capela et al. [344] tested three different feature selection algorithms and found that mean accelerations were consistently selected to distinguish between sitting, standing, and lying down, in able-bodied older adults (74±6.3 years). Similarly, Pannurat et al. [303] found that mean accelerations of the waist were ranked amongst the six most important features for classification of walking, standing, sitting, and lying down in healthy adults (average age of 67.5 years). Finally, Bijnens et al. [276] demonstrated the feasibility of using mean accelerations of the upper thigh to distinguish between standing and sitting/lying down in healthy older adults between the ages of 60-88 years. Hence, our findings underline and reconfirm the importance of the mean for distinguishing between static activities whose differences are primarily defined by posture.

Secondly, to distinguish between standing and walking, which are similar in posture, Min-X UT and RMS-X UT appeared to provide discriminative information. Despite the widespread use of the minimum [297, 298, 300, 302, 303] and RMS [297, 298, 300, 306, 330, 331] in HAR, their exact discriminative properties are ill-defined in literature. However, it is known that the RMS is sensitive to signal variability and that the minimum is sensitive to signal amplitudes. Both of these properties are distinctive for static and dynamic activities, which could explain their ability to discriminate between standing and walking.

Thirdly, the contribution of the aforementioned features in relation to recognition of transfers remains challenging to explain. The primary challenge resides in the fact that transfers always occur in between two other activities, which causes their feature values to overlap. In the feature selection study by Capela et al. [344], no generalisable features for the characterisation of transfers could be identified either. However, others did find that accelerations of the chest, waist, and upper thigh effectively captured the range of motion of transfers, and thereby enhanced recognition rates [311]. We extracted features from two similar locations: the lower back and upper thigh. We demonstrated that the six features extracted from these locations could collectively recognise transfers quite accurately with an F1-score of 0.816±0.120. Hence, since the most crucial features could not be pinpointed from a univariate standpoint, it is postulated that interactions between these features underpinned the success of the good transfer recognition rate.

Although our model generally performed well on most participants, some issues with regards to inter-person variability were observed for the detection of lying down and walking. Firstly, the HAR algorithm misclassified all instances of lying down as sitting for participant 18. Amongst participants for whom lying down was mostly correctly classified, the Mean-X LB values primarily varied between -0.1-0.1g. The Mean-X LB values of participant 18 were more deviant with values near -0.4g, which may also be observed during a backward leaning sitting posture. This could potentially explain why instances of lying down were misclassified as sitting in participant 18. There are several external factors which may have caused this deviating feature value, such as physiological differences between participants which altered the sensor orientation, or sensor displacement during execution of ADL tasks. Clinical practitioners may need to be particularly mindful of the latter, as displacements are more likely to occur during long-term monitoring in FLC [311], which may harm the HAR algorithm's ability to accurately recognise physical activities [415].

Secondly, the HAR algorithm misclassified most instances of walking as standing for participant 1. The misclassifications were postulated to be due to the fact that participant 1 showcased noticeably smaller acceleration amplitudes during walking, which decreased the discriminative properties of Min-X UT to distinguish it from standing. This may explain why walking was often mistaken for standing in participant 1. These observations entail that our HAR algorithm recognises ambulation behaviours less accurately in individuals whose walking accelerations exhibit lower amplitudes, e.g. due to slower or shuffling gait.

The limitation of our algorithm in accurately recognising slow or shuffling gait presents a challenge when applied to the hip fracture patient population. These forms of ambulation may be more prominent amongst geriatric hip fracture patients, considering that more than half of them do not regain their prefracture mobility level within the first postoperative year [32, 354]. Following hip fracture surgery, increased double support time, increased single support asymmetry, decreased cadence, and increased step length may be observed amongst patients [416]. These factors highlight the need to consider the unique gait characteristics of hip fracture patients.

Apart from deviating ambulation patterns, transfers may also look different in hip fracture patients. Compared to healthy older adults (69.4±10.9 years), patients recovering from a hip fracture (76.4±7.1 years) rely significantly more on force compensations from the contralateral side of the fractured hip to perform sit-to-stand transfers [417]. Besides force asymmetry in the lower extremities, transfer times are generally prolonged for hip fracture patients. The cohort examined in our study showcased an average sit-to-stand transfer time of 2.16 seconds. This is comparable to other studies examining healthy older adults, in which average sit-to-stand transfers times of 2.41-2.90 seconds were reported [418–421]. However, for rehabilitating hip fracture patients, the average transfer time was previously estimated at approximately 5.35 seconds [418]. In conclusion, the deviating transfer movements and prolonged transfer times highlight the potential generalisation problems that may be encountered when using our algorithm in the patient population.

Based on the aforementioned limitations, developing an algorithm in a cohort that is age-matched with the average hip fracture population may not be enough to fully optimise the algorithm's generalisation capabilities. In line with a systematic review on HAR for health research [285], we want to emphasise the importance of producing HAR algorithms in a diverse participant pool to improve the population-level impact. Previous studies have already demonstrated that diversifying the study population in terms of physical activity performance leads to HAR algorithms with better generalisation properties [317]. These advancements towards generalisation are postulated to be particularly important for monitoring systems geared towards hip fracture patients, since heterogeneity in movement patterns due to fracture type [416, 422] and recovery stage may be observed.

Despite the absence of hip fracture patients in our study cohort, we believe that the contributions of this study are still valuable. Firstly, contrary to commercially available activity trackers, we developed our HAR algorithm in a demographic age group which is representative of the hip fracture patient population. Secondly, a well-functioning monitoring system should be able to recognise physical activities across patients with a wide range of prognoses, which includes those with a more swift and better restitution of their prefracture mobility levels. Thirdly, we have pinpointed several focus areas for future

researchers to improve the generalisability of the proposed monitoring system. Hence, the HAR algorithm presented here provides the first step in the development of a continuous monitoring system which can be applied across all rehabilitating hip fracture patients.

5.5 Conclusion

This study focused on the development of a continuous monitoring system for the detection of physical activities during hip fracture rehabilitation. The activities of interest were walking, standing, sitting, lying down and transfers. Firstly, we accounted for intra-activity variation by collecting data in simulated FLC. Secondly, robustness to inter-person variability was accounted for through the use of a synthetic data set which represented the common gait characteristics across different individuals. We found that feature extraction from synthetic data had the potential to enhance classification performance on real data, which was most noticeable in the improved recognition rate of lying down and transfers. The developed monitoring system showcased good predictive abilities, using only six features extracted from a total of two accelerometers placed on the upper thigh and lower back. While the preliminary results are promising, validation in the hip fracture patient population is necessary to examine the system's true utility in clinical practice.

CHAPTER 6

Summary and Future Perspectives

The objective of this thesis was to optimise care for two distinct subgroups in the hip fracture patient population: (1) frail patients with a limited life expectancy and (2) resilient patients for whom functional recovery is feasible. To achieve the former aim, we focused on decision support for optimal treatment choices in the preoperative phase. To achieve the latter aim, we focused on the development of a continuous monitoring system which can be used to gain more insights into a patient's restitution of physical activity during rehabilitation. In this chapter, we provide a summary of the main findings of this thesis, along with recommendations for future research.

6.1 Part I: Optimal Preoperative Decision-Making

Main Findings

In Chapter 2, we started with the identification of patients who could potentially be unfit for surgery to aid prevention of surgical overtreatment. This was done by means of a systematic review and meta-analysis of preoperative predictors for early postoperative mortality. Subsequently, the identified predictors were embedded into a clinical vignette study to examine how they would affect surgeons' risk perceptions and preferences for recommending nonoperative management (Chapter 3). Through these two studies, we provided decision support for electing conservative treatment using an empirical risk perspective, and surgeons' clinical expertise.

The main findings were consistent with the national guidelines, which recommend palliative nonoperative management for patients who are at a high risk of perioperative death. Amongst the examined attributes, we found that metastatic carcinoma, severe heart failure, end-stage renal failure, and dementia increased surgeons' perceived utility of conservative treatment the most. The first three attributes in particular, were associated with a high prognostic value for postoperative 30-day mortality according to our systematic

review. Although surgeons agreed that the utility of conservative treatment was higher for patients who were at a higher risk of early mortality, we observed that surgeons' risk perceptions for the same patient cases were highly heterogeneous. Therefore, the use of objective mortality prediction models is necessary in clinical practice to streamline risk perceptions across surgeons.

However, we hypothesise that mortality prediction models alone are insufficient to identify patients who may benefit from conservative treatment. We observed that the presence of dementia substantially influenced surgeons' perceived benefit of conservative treatment, even though our meta-analysis proposed that dementia only increased mortality risk with a small-to-moderate effect size. In fact, the observed effect size was twice as large as the prior mean we specified in our hierarchical Bayesian logit model. Where our prior specification fell short, was that we did not consider the utility of conservative treatment in terms of quality of life (QoL) prospects. In clinical practice, QoL does play an important role in electing conservative treatment. Thus, it is postulated that the observed increase in utility is derived from poor postoperative QoL prognoses, on top of increased mortality risk. Based on surgeons' stated preferences, we conclude that decision support for electing nonoperative management requires more active considerations of QoL prospects.

Future Perspectives

We see several interesting directions for future research to optimise the preoperative decision-making process. Although we advocate the use of objective mortality prediction models in clinical practice, we believe that it could be valuable to move beyond mortality as the model's outcome variable. To provide more holistic decision support, future studies could develop models which predict the optimal treatment choice based on attributes which encompass both mortality risk and QoL considerations. The task of predicting the optimal treatment can be conceptualised as a choice modelling problem, in which we aim to elect the treatment which maximises a patient's utility. From this problem formulation follows that a health preference research framework is suitable for data collection and modelling. Based on the lessons learnt, we propose several recommendations to accomplish this.

As described in Chapter 3, surgeons repeatedly expressed that they missed information on patients' personal preferences while assessing the vignettes. Some surgeons also stated that the opinions of geriatricians and anaesthesiologists could have helped them in providing better treatment recommendations. Omission of views and values of core actors in the decision context could harm the external validity of the stated optimal treatment choice. Therefore, we provide two recommendations to safeguard the external validity of the prediction model. Firstly, the attributes presented in Chapter 3 could be updated by incorporating important patient reported outcome measures (PROMs) into the vignettes. We believe that PROMs can provide a more holistic perspective on which QoL considerations are important to patients during the final phases of their lives. Hence, we encourage researchers to actively involve conservatively treated patients and their caregivers into the research process. Secondly, preference elicitation could be performed through a multidisciplinary consultation, to ensure that the expertise of the complete medical team is reflected in the optimal treatment choice.

6.2 Part II: Optimal Monitoring during Rehabilitation

Main Findings

In Chapter 4, we provided a critical perspective on common practices in the development of human activity recognition (HAR) algorithms based on wearable accelerometers. We observed that common practices were often transferred from one setting to another, without keeping the specifics of the study cohort, experimental setup, and performed activities in mind. It was postulated that such practices would be particularly harmful for feature selection, and thus for the system's performance, since the informativeness of features depends strongly on the aforementioned specifics. Through synthesis of evidence found in literature, we provided concrete recommendations to practitioners for various (algorithmic) design choices for the development of ambulatory monitoring systems. Amongst others, these included recommendations for choosing optimal accelerometer placements, adequate signal segmentation techniques, and effective feature selection methods. We used these best practices to inform the design of our monitoring system described in Chapter 5.

In Chapter 5, we developed a monitoring system which could reliably detect physical activities and important functional milestones for rehabilitating hip fracture patients. We showed that it was feasible to develop a high-performing activity classification algorithm based on machine learning, using a total of two accelerometers placed on the upper thigh and lower back. Based on leave-one-subject-out cross-validation, we estimated that the system had a low prediction error for physical activity classifications in healthy adults aged 80 years or older, as the mean and standard deviations of the F1-scores were 0.896 ± 0.100 for walking, 0.927 ± 0.039 for standing, 0.997 ± 0.004 for sitting, 0.937 ± 0.202 for lying down and 0.816 ± 0.120 for transfers.

To develop this system, we introduced a novel feature selection pipeline to enhance the robustness and generalisability of the selected features. First, we created a synthetic data set based on common gait patterns across different individuals, to improve robustness to inter-person variability. Subsequently, we performed feature selection on the synthetic data set. We controlled the robustness of the feature selection process in two-fold. Firstly, we developed a stable heterogeneous feature selection ensemble (HFSE) to ensure that the selected features would be robust to different heuristics for determining feature importance. Secondly, we reduced the likelihood of including features that were selected by chance, by imposing selection stability across different partitions of the data as an additional criterion. We found that the proposed pipeline was capable of reducing the initial candidate set of 62 features, down to solely six highly informative features. Compared to strictly using real data for model building, we observed that synthetic data manipulation yielded a slight improvement in overall classification performance, which was most noticeable in the recognition rates of transfers and lying down.

Future Perspectives

First and foremost, we believe that it is necessary to validate the developed algorithm in the hip fracture patient population to assess its generalisation capabilities. As argued in Chapter 5, we foresee several challenges in the recognition of hip fracture patients'

ambulation and transfer behaviours. Specifically, our system may underestimate the time spent in ambulation for patients with slow or shuffling gait, and it may fail to detect slow transfers. Therefore, we recommend future researchers to augment the data collected in our study with physical activity data of hip fracture patients. We strongly believe that this approach will contribute to improved generalisability the most, to ultimately make the monitoring system useful for all rehabilitating hip fracture patients.

Besides focusing on generalisability, we believe that there is still room for improvement in the recognition rate of transfers. The current monitoring system still faces some challenges in distinguishing between dynamic sitting activities, such as adjustments of sitting postures, from transfers. Since transfers were heavily underrepresented in our data set, it is possible that the HFSE was biased towards selecting features which were informative for recognising the majority class, i.e. sitting. Therefore, it could be interesting to examine whether better features for recognition of transfers could be identified by applying the HFSE to a synthetic data set without class imbalance.

6.3 Final Remarks

This thesis contributes to our understanding on how to optimise care for specific subgroups in the hip fracture patient population. We hope that our findings serve as a reference framework for future researchers, who continue to work on improving preoperative decision-making and monitoring during postoperative rehabilitation. Ultimately, we hope that the contributions described in this thesis may benefit hip fracture patients through improvements in the quality of care in the future.

Bibliography

[1] X. L. Griffin, N. Parsons, J. Achten, M. Fernandez, and M. L. Costa, "Recovery of health-related quality of life in a United Kingdom hip fracture population: the Warwick Hip Trauma Evaluation-a prospective cohort study," *The bone & joint journal*, vol. 97, no. 3, pp. 372–382, 2015. DOI: 10.1302/0301-620X.97B3.35738.

[2] S. Haleem, L. Lutchman, R. Mayahi, J. Grice, and M. Parker, "Mortality following hip fracture: Trends and geographical variations over the last 40 years," *Injury*, vol. 39, no. 10, pp. 1157–1163, Oct. 2008, ISSN: 00201383. DOI: 10.1016/j.injury.2008.03.022.

[3] K. Rapp, G. Büchele, K. Dreinhöfer, B. Bücking, C. Becker, and P. Benzinger, "Epidemiology of hip fractures: Systematic literature review of German data and an overview of the international literature," *Zeitschrift für Gerontologie und Geriatrie*, vol. 52, no. 1, pp. 10–16, 2019. DOI: 10.1007/s00391-018-1382-z.

[4] Z. Liu, J. Zhang, K. He, Y. Zhang, and Y. Zhang, "Optimized clinical practice for superaged patients with hip fracture: significance of damage control and enhanced recovery program," *Burns & trauma*, vol. 7, 2019. DOI: 10.1186/s41038-019-0159-y.

[5] B. Gullberg, O. Johnell, and J. A. Kanis, "World-wide projections for hip fracture," *Osteoporosis international*, vol. 7, no. 5, pp. 407–413, 1997. DOI: 10.1007/pl00004148.

[6] P. Kannus, J. Parkkari, H. Sievänen, A. Heinonen, I. Vuori, and M. Järvinen, "Epidemiology of hip fractures," *Bone*, vol. 18, no. 1, S57–S63, 1996. DOI: 10.1016/8756-3282(95)00381-9.

[7] R. Marks, "Hip fracture epidemiological trends, outcomes, and risk factors, 1970–2009," *International journal of general medicine*, vol. 3, p. 1, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2866546/pdf/ijgm-3-001.pdf (visited on 06/06/2021).

[8] J. D. Penrod, A. Litke, W. G. Hawkes, J. Magaziner, K. J. Koval, J. T. Doucette, S. B. Silberzweig, and A. L. Siu, "Heterogeneity in Hip Fracture Patients: Age, Functional Status, and Comorbidity," *Journal of the American Geriatrics Society*, vol. 55, no. 3, pp. 407–413, 2007, ISSN: 1532-5415. DOI: 10.1111/j.1532-5415.2007.01078.x.

[9] M. Parker and A. Johansen, "Hip fracture," *Bmj*, vol. 333, no. 7557, pp. 27–30, 2006. DOI: 10.1136/bmj.333.7557.27.

[10] M. Fransen, M. Woodward, R. Norton, E. Robinson, M. Butler, and A. J. Campbell, "Excess mortality or institutionalization after hip fracture: men are at greater risk

than women," *Journal of the American Geriatrics Society*, vol. 50, no. 4, pp. 685–690, Apr. 2002, ISSN: 0002-8614. DOI: 10.1046/j.1532-5415.2002.50163.x.

[11] H. E. Whitson, W. Duan-Porter, K. E. Schmader, M. C. Morey, H. J. Cohen, and C. S. Colón-Emeric, "Physical Resilience in Older Adults: Systematic Review and Development of an Emerging Construct," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 71, no. 4, pp. 489–495, Apr. 2016, ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glv202.

[12] J. L. Wolff, B. Starfield, and G. Anderson, "Prevalence, Expenditures, and Complications of Multiple Chronic Conditions in the Elderly," *Archives of Internal Medicine*, vol. 162, no. 20, pp. 2269–2276, Nov. 2002, ISSN: 0003-9926. DOI: 10.1001/archinte.162.20.2269.

[13] E. R. Flikweert, K. W. Wendt, R. L. Diercks, G. J. Izaks, D. Landsheer, M. Stevens, and I. H. F. Reininga, "Complications after hip fracture surgery: are they preventable?" *European Journal of Trauma and Emergency Surgery*, vol. 44, no. 4, pp. 573–580, Aug. 2018, ISSN: 1863-9933, 1863-9941. DOI: 10.1007/s00068-017-0826-2.

[14] P. J. Belmont, J. G. E'Stephan, D. Romano, J. O. Bader, K. J. Nelson, and A. J. Schoenfeld, "Risk factors for complications and in-hospital mortality following hip fractures: a study using the National Trauma Data Bank," *Archives of orthopaedic and trauma surgery*, vol. 134, no. 5, pp. 597–604, 2014. DOI: 10.1007/s00402-014-1959-y.

[15] P. Carpintero, J. R. Caeiro, R. Carpintero, A. Morales, S. Silva, and M. Mesa, "Complications of hip fractures: A review," *World Journal of Orthopedics*, vol. 5, no. 4, pp. 402–411, Sep. 2014, ISSN: 2218-5836. DOI: 10.5312/wjo.v5.i4.402.

[16] Y. Luo, Y. Jiang, H. Xu, H. Lyu, L. Zhang, P. Yin, and P. Tang, "Risk of post-operative cardiovascular event in elderly patients with pre-existing cardiovascular disease who are undergoing hip fracture surgery," *International Orthopaedics*, vol. 45, no. 12, pp. 3045–3053, Dec. 2021, ISSN: 0341-2695, 1432-5195. DOI: 10.1007/s00264-021-05227-7.

[17] M. W. Cullen, R. E. Gullerud, D. R. Larson, L. J. Melton III, and J. M. Huddleston, "Impact of heart failure on hip fracture outcomes: A population-based study," *Journal of Hospital Medicine*, vol. 6, no. 9, pp. 507–512, 2011, ISSN: 1553-5606. DOI: 10.1002/jhm.918.

[18] C. J. Porter, I. K. Moppett, I. Juurlink, J. Nightingale, C. G. Moran, and M. A. J. Devonald, "Acute and chronic kidney disease in elderly patients with hip fracture: prevalence, risk factors and outcome with development and validation of a risk prediction model for acute kidney injury," *BMC Nephrology*, vol. 18, no. 1, p. 20, Dec. 2017, ISSN: 1471-2369. DOI: 10.1186/s12882-017-0437-5.

[19] M. Bhandari and M. Swiontkowski, "Management of Acute Hip Fracture," *New England Journal of Medicine*, vol. 377, no. 21, pp. 2053–2062, Nov. 2017, Publisher: Massachusetts Medical Society, ISSN: 0028-4793. DOI: 10.1056/NEJMcp1611090.

[20] M. J. Parker and C. R. Palmer, "A new mobility score for predicting mortality after hip fracture," *The Journal of bone and joint surgery. British volume*, vol. 75, no. 5, pp. 797–798, 1993. DOI: 10.1302/0301-620X.75B5.8376443.

[21] I. Spronk, S. A. I. Loggers, P. Joosse, H. C. Willems, R. Van Balen, T. Gosens, K. J. Ponsen, J. Steens, L. C. P. (Van de Ree, R. G. Zuurmond, M. H. J. Verhofstad, E. M. M. Van Lieshout, and S. Polinder, "Shared decision-making for the treatment of proximal femoral fractures in frail institutionalised older patients: healthcare providers' perceived barriers and facilitators," *Age and Ageing*, vol. 51, no. 8, afac174, Aug. 2022, ISSN: 0002-0729, 1468-2834. DOI: 10.1093/ageing/afac174.

[22] L. Moja, A. Piatti, V. Pecoraro, C. Ricci, G. Virgili, G. Salanti, L. Germagnoli, A. Liberati, and G. Banfi, "Timing Matters in Hip Fracture Surgery: Patients Operated within 48 Hours Have Better Outcomes. A Meta-Analysis and Meta-Regression of over 190,000 Patients," *PLoS ONE*, vol. 7, no. 10, R. W. Scherer, Ed., e46175, Oct. 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046175.

[23] S. A. I. Loggers, E. M. M. Van Lieshout, P. Joosse, M. H. J. Verhofstad, and H. C. Willems, "Prognosis of nonoperative treatment in elderly patients with a hip fracture: A systematic review and meta-analysis," *Injury*, vol. 51, no. 11, pp. 2407–2413, Nov. 2020, ISSN: 1879-0267. DOI: 10.1016/j.injury.2020.08.027.

[24] P. McNamara and K. Sharma, "Surgery or palliation for hip fractures in patients with advanced malignancy?" *Age and Ageing*, vol. 26, no. 6, pp. 471–474, Nov. 1997, ISSN: 0002-0729. DOI: 10.1093/ageing/26.6.471.

[25] L. K. Cannada, S. C. Mears, and C. Quatman, "Clinical Faceoff: When Should Patients 65 Years of Age and Older Have Surgery for Hip Fractures, and When is it a Bad Idea?" *Clinical Orthopaedics and Related Research*, vol. 479, no. 1, pp. 24–27, Jan. 2021, ISSN: 1528-1132. DOI: 10.1097/CORR.0000000000001596.

[26] Federatie Medisch Specialisten, *Richtlijn proximale femurfracturen*. [Online]. Available: https://richtlijnendatabase.nl/richtlijn/proximale_femurfracturen/proximale_femurfracturen_-_startpagina.html (visited on 04/20/2022).

[27] D. Stow, G. Spiers, F. E. Matthews, and B. Hanratty, "What is the evidence that people with frailty have needs for palliative care at the end of life? A systematic review and narrative synthesis," *Palliative Medicine*, vol. 33, no. 4, pp. 399–414, Apr. 2019, Publisher: SAGE Publications Ltd STM, ISSN: 0269-2163. DOI: 10.1177/0269216319828650.

[28] S. A. I. Loggers, H. C. Willems, R. Van Balen, T. Gosens, S. Polinder, K. J. Ponsen, C. L. P. Van de Ree, J. Steens, M. H. J. Verhofstad, R. G. Zuurmond, E. M. M. Van Lieshout, P. Joosse, and F.-H. S. Group, "Evaluation of Quality of Life After Nonoperative or Operative Management of Proximal Femoral Fractures in Frail Institutionalized Patients: The FRAIL-HIP Study," *JAMA Surgery*, Mar. 2022, ISSN: 2168-6254. DOI: 10.1001/jamasurg.2022.0089.

[29] F. J. Amarilla-Donoso, F. López-Espuela, R. Roncero-Martín, O. Leal-Hernandez, L. M. Puerto-Parejo, I. Aliaga-Vera, R. Toribio-Felipe, and J. M. Lavado-García, "Quality of life in elderly people after a hip fracture: a prospective study," *Health and Quality of Life Outcomes*, vol. 18, no. 1, p. 71, Mar. 2020, ISSN: 1477-7525. DOI: 10.1186/s12955-020-01314-2.

[30] J.-E. Gjertsen, V. Baste, J. M. Fevang, O. Furnes, and L. B. Engesæter, "Quality of life following hip fractures: results from the Norwegian hip fracture register," *BMC*

musculoskeletal disorders, vol. 17, no. 1, pp. 1–8, 2016. DOI: 10.1186/s12891-016-1111-y.

[31] R. Brooks, R. Rabin, and F. Charro, Eds., *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective*, 1st ed. Springer Dordrecht, 2003, ISBN: 978-94-017-0233-1. [Online]. Available: https://link.springer.com/book/10.1007/978-94-017-0233-1 (visited on 07/31/2022).

[32] M. Bertram, R. Norman, L. Kemp, and T. Vos, "Review of the long-term disability associated with hip fractures," *Injury Prevention*, vol. 17, no. 6, pp. 365–370, Dec. 2011, ISSN: 1353-8047, 1475-5785. DOI: 10.1136/ip.2010.029579.

[33] G. Salkeld, I. D. Cameron, R. G. Cumming, S. Easter, J. Seymour, S. E. Kurrle, and S. Quine, "Quality of life related to fear of falling and hip fracture in older women: a time trade off study," *BMJ (Clinical research ed.)*, vol. 320, no. 7231, pp. 341–346, Feb. 2000, ISSN: 0959-8138. DOI: 10.1136/bmj.320.7231.341.

[34] S. M. Robinson, B. Ní Bhuachalla, B. Ní Mhaille, P. E. Cotter, M. O'Connor, and S. T. O'Keeffe, "Home, please: A conjoint analysis of patient preferences after a bad hip fracture," *Geriatrics & Gerontology International*, vol. 15, no. 10, pp. 1165–1170, Oct. 2015, ISSN: 1447-0594. DOI: 10.1111/ggi.12415.

[35] D. van Dartel, J. H. Hegeman, and M. M. R. Vollenbroek-Hutten, "Feasibility and Usability of Wearable Devices for Ambulatory Monitoring of the Rehabilitation Process of Older Patients after Hip Fracture Surgery," in *Proceedings of the 18th International Conference on Wireless Networks and Mobile Systems - WINSYS*, INSTICC, SciTePress, 2021, pp. 59–66. DOI: 10.5220/0010522500590066.

[36] N. Radosavljevic, D. Nikolic, M. Lazovic, and A. Jeremic, "Hip fractures in a geriatric population-rehabilitation based on patients needs," *Aging and disease*, vol. 5, no. 3, p. 177, 2014. DOI: 10.14336/AD.2014.0500177.

[37] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, pp. 1–17, 2012. DOI: 10.1186/1743-0003-9-21.

[38] P. Benzinger, U. Lindemann, C. Becker, K. Aminian, M. Jamour, and S. Flick, "Geriatric rehabilitation after hip fracture: Role of body-fixed sensor measurements of physical activity," *Zeitschrift für Gerontologie und Geriatrie*, vol. 47, no. 3, pp. 236–242, Apr. 2014, ISSN: 0948-6704, 1435-1269. DOI: 10.1007/s00391-013-0477-9.

[39] K. Taraldsen, O. Sletvold, P. Thingstad, I. Saltvedt, M. H. Granat, S. Lydersen, and J. L. Helbostad, "Physical Behavior and Function Early After Hip Fracture Surgery in Patients Receiving Comprehensive Geriatric Care or Orthopedic Care–A Randomized Controlled Trial," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 69A, no. 3, pp. 338–345, Mar. 2014, ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glt097.

[40] M. Fitzgerald, C. Blake, D. Askin, J. Quinlan, T. Coughlan, and C. Cunningham, "Mobility one week after a hip fracture – can it be predicted?" *International Journal of Orthopaedic and Trauma Nursing*, vol. 29, pp. 3–9, May 2018, ISSN: 1878-1241. DOI: 10.1016/j.ijotn.2017.11.001.

[41] S. J. Davenport, M. Arnold, C. Hua, A. Schenck, S. Batten, and N. F. Taylor, "Physical Activity Levels During Acute Inpatient Admission After Hip Fracture are

Very Low," *Physiotherapy Research International*, vol. 20, no. 3, pp. 174–181, 2015, ISSN: 1471-2865. DOI: 10.1002/pri.1616.

[42] E. Willems, J. Visschedijk, R. van Balen, and W. Achterberg, "Physical Activity, Physical Function and Fear of Falling After Hip Fracture," *Orthopedic Research & Physiotherapy*, vol. 3, no. 1, pp. 1–6, Jul. 2017, ISSN: 23812052. DOI: 10.24966/ORP-2052/100031.

[43] E. A. Eastwood, J. Magaziner, J. Wang, S. B. Silberzweig, E. L. Hannan, E. Strauss, and A. L. Siu, "Patients with Hip Fracture: Subgroups and Their Outcomes," *Journal of the American Geriatrics Society*, vol. 50, no. 7, pp. 1240–1249, 2002, ISSN: 1532-5415. DOI: 10.1046/j.1532-5415.2002.50311.x.

[44] J.-P. Michel, P. Hoffmeyer, C. Klopfenstein, M. Bruchez, B. Grab, and C. L. d'Epinay, "Prognosis of Functional Recovery 1 Year After Hip Fracture: Typical Patient Profiles Through Cluster Analysis," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 55, no. 9, pp. M508–M515, Sep. 2000, ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/55.9.M508.

[45] T. Smith, K. Pelpola, M. Ball, A. Ong, and P. K. Myint, "Pre-operative indicators for mortality following hip fracture surgery: a systematic review and meta-analysis," *Age and Ageing*, vol. 43, no. 4, pp. 464–471, Jul. 2014, ISSN: 0002-0729. DOI: 10.1093/ageing/afu065.

[46] W. Chang, H. Lv, C. Feng, P. Yuwen, N. Wei, W. Chen, and Y. Zhang, "Preventable risk factors of mortality after hip fracture surgery: Systematic review and meta-analysis," *International Journal of Surgery*, vol. 52, pp. 320–328, Apr. 2018, ISSN: 1743-9191. DOI: 10.1016/j.ijsu.2018.02.061.

[47] F. Hu, C. Jiang, J. Shen, P. Tang, and Y. Wang, "Preoperative predictors for mortality following hip fracture surgery: a systematic review and meta-analysis," *Injury*, vol. 43, no. 6, pp. 676–685, 2012. DOI: 10.1016/j.injury.2011.05.017.

[48] D. Stacey, F. Légaré, K. Lewis, M. J. Barry, C. L. Bennett, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson, *et al.*, "Decision aids for people facing health treatment or screening decisions," *Cochrane database of systematic reviews*, no. 4, 2017. DOI: 10.1002/14651858.CD001431.pub5.

[49] E. A. G. Joosten, L. DeFuentes-Merillas, G. H. De Weert, T. Sensky, C. P. F. Van Der Staak, and C. A. J. de Jong, "Systematic review of the effects of shared decision-making on patient satisfaction, treatment adherence and health status," *Psychotherapy and psychosomatics*, vol. 77, no. 4, pp. 219–226, 2008. DOI: 10.1159/000126073.

[50] A. M. Stiggelbout, T. Van der Weijden, M. P. De Wit, D. Frosch, F. Légaré, V. M. Montori, L. Trevena, and G. Elwyn, "Shared decision making: really putting patients at the centre of healthcare," *Bmj*, vol. 344, 2012. DOI: 10.1136/bmj.e256.

[51] A.-B. Haidich, "Meta-analysis in medical research," *Hippokratia*, vol. 14, no. Suppl 1, p. 29, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/pdf/hippokratia-14-29.pdf (visited on 09/17/2021).

[52] R. M. Turner, J. Davey, M. J. Clarke, S. G. Thompson, and J. P. T. Higgins, "Predicting the extent of heterogeneity in meta-analysis, using empirical data

from the Cochrane Database of Systematic Reviews," *International journal of epidemiology*, vol. 41, no. 3, pp. 818–827, 2012. DOI: 10.1093/ije/dys041.

[53] E. Kontopantelis, D. A. Springate, and D. Reeves, "A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses," *PloS one*, vol. 8, no. 7, e69930, 2013. DOI: 10.1371/journal.pone.0069930.

[54] Y. Chung, S. Rabe-Hesketh, and I.-H. Choi, "Avoiding zero between-study variance estimates in random-effects meta-analysis," *Statistics in medicine*, vol. 32, no. 23, pp. 4071–4089, 2013. DOI: 10.1002/sim.5821.

[55] D. R. Williams, P. Rast, and P.-C. Bürkner, "Bayesian meta-analysis with weakly informative prior distributions," 2018. DOI: 10.31234/osf.io/7tbrm.

[56] K. M. Rhodes, R. M. Turner, and J. P. T. Higgins, "Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data," *Journal of clinical epidemiology*, vol. 68, no. 1, pp. 52–60, 2015. DOI: 10.1016/j.jclinepi.2014.08.012.

[57] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to meta-analysis*. John Wiley & Sons, 2009. DOI: 10.1002/9780470743386.

[58] C. Röver, G. Knapp, and T. Friede, "Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies," *BMC medical research methodology*, vol. 15, no. 1, pp. 1–7, 2015. DOI: 10.1186/s12874-015-0091-1.

[59] F. Foroutan, G. Guyatt, V. Zuk, P. O. Vandvik, A. C. Alba, R. Mustafa, R. Vernooij, I. Arevalo-Rodriguez, Z. Munn, P. Roshanov, *et al.*, "GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks," *Journal of clinical epidemiology*, vol. 121, pp. 62–70, 2020. DOI: 10.1016/j.jclinepi.2019.12.023.

[60] A. Iorio, F. A. Spencer, M. Falavigna, C. Alba, E. Lang, B. Burnand, T. McGinn, J. Hayden, K. Williams, B. Shea, R. Wolff, T. Kujpers, P. Perel, P. O. Vandvik, P. Glasziou, H. Schunemann, and G. Guyatt, "Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients," *BMJ*, vol. 350, no. mar16 7, h870–h870, Mar. 2015, ISSN: 1756-1833. DOI: 10.1136/bmj.h870.

[61] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *Bmj*, vol. 372, 2021. DOI: 10.1186/s13643-021-01626-4.

[62] J. A. Hayden, D. A. van der Windt, J. L. Cartwright, P. Côté, and C. Bombardier, "Assessing bias in studies of prognostic factors," *Annals of internal medicine*, vol. 158, no. 4, pp. 280–286, 2013. DOI: 10.7326/0003-4819-158-4-201302190-00009.

[63] A. Huguet, J. A. Hayden, J. Stinson, P. J. McGrath, C. T. Chambers, M. E. Tougas, and L. Wozney, "Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework," *Systematic Reviews*, vol. 2, p. 71, Sep. 2013, ISSN: 2046-4053. DOI: 10.1186/2046-4053-2-71.

[64] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Controlled clinical trials*, vol. 7, no. 3, pp. 177–188, 1986. DOI: 10.1016/0197-2456(86)90046-2.

[65] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," *Research synthesis methods*, vol. 1, no. 2, pp. 97–111, 2010. DOI: 10.1002/jrsm.12.

[66] J. P. Higgins, S. G. Thompson, and D. J. Spiegelhalter, "A re-evaluation of random-effects meta-analysis," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 137–159, 2009. DOI: 10.1111/j.1467-985X.2008.00552.x.

[67] S. Duval and R. Tweedie, "A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis," *Journal of the american statistical association*, vol. 95, no. 449, pp. 89–98, 2000. DOI: 10.1080/01621459.2000.10473905.

[68] ——, "Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis," *Biometrics*, vol. 56, no. 2, pp. 455–463, 2000. DOI: 10.1111/j.0006-341x.2000.00455.x.

[69] G. Knapp and J. Hartung, "Improved tests for a random effects meta-regression with a single covariate," *Statistics in medicine*, vol. 22, no. 17, pp. 2693–2710, 2003. DOI: 10.1002/sim.1482.

[70] M. Harrer, P. Cuijpers, F. T. A, and D. D. Ebert, *Doing Meta-Analysis With R: A Hands-On Guide*, 1st. Boca Raton, FL and London: Chapman & Hall/CRC Press, 2021, ISBN: 9780367610074. DOI: 10.1201/9781003107347.

[71] W. Viechtbauer, "Conducting Meta-Analyses in R with the metafor Package," *Journal of Statistical Software*, vol. 36, pp. 1–48, Aug. 2010, ISSN: 1548-7660. DOI: 10.18637/jss.v036.i03.

[72] P.-C. Bürkner, "brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software, Articles*, vol. 80, no. 1, 2017. DOI: 10.18637/jss.v080.i01.

[73] L. A. McGuinness and J. P. T. Higgins, "Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments," *Research Synthesis Methods*, vol. 12, no. 1, pp. 55–61, 2021, ISSN: 1759-2887. DOI: 10.1002/jrsm.1411.

[74] C. de Luise, M. Brimacombe, L. Pedersen, and H. T. Sørensen, "Comorbidity and mortality following hip fracture: a population-based cohort study," *Aging clinical and experimental research*, vol. 20, no. 5, pp. 412–418, 2008. DOI: 10.1007/BF03325146.

[75] F. S. Würdemann, J. A. Wilschut, and J. H. Hegeman, "Eindverslag SKMS Project Doorontwikkeling DHFA," Dutch Institute for Clinical Auditing, Tech. Rep. 1.1, 2021. [Online]. Available: https://dica.nl/media/2605/Verslag%5C%20SKMS%5C%20-%5C%20Doorontwikkeling%5C%20DHFA%5C%20(Indicatoren%5C%20Taskforce)_Def.pdf (visited on 02/12/2022).

[76] Y. Cao, M. P. Forssten, A. Mohammad Ismail, T. Borg, I. Ioannidis, S. Montgomery, and S. Mohseni, "Predictive Values of Preoperative Characteristics for 30-Day Mortality in Traumatic Hip Fracture Patients," *Journal of personalized medicine*, vol. 11, no. 5, p. 353, 2021. DOI: 10.3390/jpm11050353.

[77] C. Forni, D. Gazineo, F. D'Alessandro, A. Fiorani, M. Morri, T. Sabattini, E. Ambrosi, and P. Chiari, "Predictive factors for thirty day mortality in geriatric

patients with hip fractures: a prospective study," *International orthopaedics*, vol. 43, no. 2, pp. 275–281, 2019. DOI: 10.1007/s00264-018-4057-x.

[78] J. Mayordomo-Cava, L. Abásolo, N. Montero-Fernandez, J. Ortiz-Alonso, M. Vidán-Astiz, and J. A. Serra-Rexach, "Hip Fracture in nonagenarians: characteristics and factors related to 30-day mortality in 1177 patients," *The Journal of arthroplasty*, vol. 35, no. 5, pp. 1186–1193, 2020. DOI: 10.1016/j.arth.2019.12.044.

[79] N. Morrissey, E. Iliopoulos, A. W. Osmani, and K. Newman, "Neck of femur fractures in the elderly: does every hour to surgery count?" *Injury*, vol. 48, no. 6, pp. 1155–1158, 2017. DOI: 10.1016/j.injury.2017.03.007.

[80] L. M. G. Nijland, J. Karres, A. E. Simons, J. M. Ultee, G. M. M. J. Kerkhoffs, and B. C. Vrouenraets, "The weekend effect for hip fracture surgery," *Injury*, vol. 48, no. 7, pp. 1536–1541, 2017. DOI: 10.1016/j.injury.2017.05.017.

[81] C. Pang, A. Aqil, A. Mannan, G. Thomas, and F. S. Hossain, "Hip fracture patients admitted to hospital on weekends are not at increased risk of 30-day mortality as compared with weekdays," *Journal of Orthopaedics and Traumatology*, vol. 21, no. 1, pp. 1–6, 2020. DOI: 10.1186/s10195-020-00558-4.

[82] C. J. Thomas, R. P. Smith, C. E. Uzoigwe, and J. R. Braybrooke, "The weekend effect: short-term mortality following admission with a hip fracture," *The bone & joint journal*, vol. 96, no. 3, pp. 373–378, 2014. DOI: 10.1302/0301-620X.96B3.33118.

[83] C. L. P. van de Ree, T. Gosens, A. H. van der Veen, C. J. M. Oosterbos, M. W. Heymans, and M. A. C. de Jongh, "Development and validation of the Brabant Hip Fracture Score for 30-day and 1-year mortality," *Hip International*, vol. 30, no. 3, pp. 354–362, 2020. DOI: 10.1177/1120700019836962.

[84] H. C. Rae, I. A. Harris, L. McEvoy, and T. Todorova, "Delay to Surgery and Mortality After Hip Fracture," *ANZ Journal of Surgery*, vol. 77, no. 10, pp. 889–891, 2007, ISSN: 1445-2197. DOI: 10.1111/j.1445-2197.2007.04267.x.

[85] B. D. Chatterton, T. S. Moores, S. Ahmad, A. Cattell, and P. J. Roberts, "Cause of death and factors associated with early in-hospital mortality after hip fracture," *The bone & joint journal*, vol. 97, no. 2, pp. 246–251, 2015. DOI: 10.1302/0301-620X.97B2.35248.

[86] N. B. Foss and H. Kehlet, "Short-term mortality in hip fracture patients admitted during weekends and holidays," *BJA: British Journal of Anaesthesia*, vol. 96, no. 4, pp. 450–454, 2006. DOI: 10.1093/bja/ael012.

[87] W. S. Nijmeijer, E. C. Folbert, M. Vermeer, J. P. Slaets, and J. H. Hegeman, "Prediction of early mortality following hip fracture surgery in frail elderly: The Almelo Hip Fracture Score (AHFS)," *Injury*, vol. 47, no. 10, pp. 2138–2143, 2016. DOI: 10.1016/j.injury.2016.07.022.

[88] H. J. Schuijt, J. Bos, D. P. J. Smeeing, O. Geraghty, and D. van der Velde, "Predictors of 30-day mortality in orthogeriatric fracture patients aged 85 years or above admitted from the emergency department," *European Journal of Trauma and Emergency Surgery*, vol. 47, no. 3, pp. 817–823, 2021. DOI: 10.1007/s00068-019-01278-z.

[89] D. Norring-Agerskov, C. Madsen, L. Bathum, O. B. Pedersen, J. B. Lauritzen, N. Jørgensen, and H. Jørgensen, "History of cardiovascular disease and cardiovascular

biomarkers are associated with 30-day mortality in patients with hip fracture," *Osteoporosis International*, vol. 30, no. 9, pp. 1767–1778, 2019. DOI: 10.1007/s00198-019-05056-w.

[90] J. Karres, N. Kieviet, J.-P. Eerenberg, and B. C. Vrouenraets, "Predicting early mortality after hip fracture surgery: the hip fracture estimator of mortality Amsterdam," *Journal of orthopaedic trauma*, vol. 32, no. 1, pp. 27–33, 2018. DOI: 10.1097/BOT.0000000000001025.

[91] A. Bottle and P. Aylin, "Mortality associated with delay in operation after hip fracture: observational study," *Bmj*, vol. 332, no. 7547, pp. 947–951, 2006. DOI: 10.1136/bmj.38790.468519.55.

[92] M. Faizi, A. J. Farrier, M. Venkatesan, C. Thomas, C. E. Uzoigwe, S. Balasubramanian, and R. P. Smith, "Is body temperature an independent predictor of mortality in hip fracture patients?" *Injury*, vol. 45, no. 12, pp. 1942–1945, 2014. DOI: 10.1016/j.injury.2014.09.024.

[93] M. J. Maxwell, C. Moran, and I. K. Moppett, "Development and validation of a preoperative scoring system to predict 30 day mortality in patients undergoing hip fracture surgery," *British journal of anaesthesia*, vol. 101, no. 4, pp. 511–517, 2008. DOI: 10.1093/bja/aen236.

[94] A. Lizaur-Utrilla, B. Gonzalez-Navarro, M. F. Vizcaya-Moreno, and F. A. Lopez-Prats, "Altered seric levels of albumin, sodium and parathyroid hormone may predict early mortality following hip fracture surgery in elderly," *International orthopaedics*, vol. 43, no. 12, pp. 2825–2829, 2019. DOI: 10.1007/s00264-019-04368-0.

[95] H. C. Chiu, C. M. Chen, T. Y. Su, C. H. Chen, H. M. Hsieh, C. P. Hsieh, and D. L. Shen, "Dementia predicted one-year mortality for patients with first hip fracture: a population-based study," *Bone Joint J*, vol. 100, no. 9, pp. 1220–1226, 2018. DOI: 10.1302/0301-620X.100B9.BJJ-2017-1342.R1.

[96] J. D. Petersen, V. D. Siersma, S. Wehberg, C. T. Nielsen, B. Viberg, and F. B. Waldorff, "Clinical management of hip fractures in elderly patients with dementia and postoperative 30-day mortality: A population-based cohort study," *Brain and Behavior*, vol. 10, no. 11, e01823, 2020. DOI: 10.1002/brb3.1823.

[97] A. Franzo, C. Francescutti, and G. Simon, "Risk factors correlated with post-operative mortality for hip fracture surgery in the elderly: a population-based approach," *European journal of epidemiology*, vol. 20, no. 12, pp. 985–991, 2005. DOI: 10.1007/s10654-005-4280-9.

[98] C. L. P. van de Ree, M. A. C. De Jongh, C. M. M. Peeters, L. de Munter, J. A. Roukema, and T. Gosens, "Hip Fractures in Elderly People: Surgery or No Surgery? A Systematic Review and Meta-Analysis," *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 8, no. 3, pp. 173–180, Sep. 2017, ISSN: 2151-4585. DOI: 10.1177/2151458517713821.

[99] L. L. Kirkland, D. T. Kashiwagi, M. C. Burton, S. Cha, and P. Varkey, "The Charlson Comorbidity Index Score as a predictor of 30-day mortality after hip fracture surgery," *American Journal of Medical Quality*, vol. 26, no. 6, pp. 461–467, 2011. DOI: 10.1177/1062860611402188.

[100] A. W. Ireland, P. J. Kelly, and R. G. Cumming, "Risk factor profiles for early and delayed mortality after hip fracture: Analyses of linked Australian Department of

Veterans' Affairs databases," *Injury*, vol. 46, no. 6, pp. 1028–1035, 2015. DOI: 10.1016/j.injury.2015.03.006.

[101] J. J. W. Roche, R. T. Wenn, O. Sahota, and C. G. Moran, "Effect of comorbidities and postoperative complications on mortality after hip fracture in elderly people: prospective observational cohort study," *Bmj*, vol. 331, no. 7529, p. 1374, 2005. DOI: 10.1136/bmj.38643.663843.55.

[102] H. Q. Sheikh, F. S. Hossain, A. Aqil, B. Akinbamijo, V. Mushtaq, and H. Kapoor, "A comprehensive analysis of the causes and predictors of 30-day mortality following hip fracture surgery," *Clinics in orthopedic surgery*, vol. 9, no. 1, pp. 10–18, 2017. DOI: 10.4055/cios.2017.9.1.10.

[103] Z. Wang, X. Chen, L. Yang, H. Wang, W. Jiang, and Y. Liu, "A new preoperative risk score for predicting mortality of elderly hip fracture patients: an external validation study," *Aging Clinical and Experimental Research*, pp. 1–9, 2021. DOI: 10.1007/s40520-021-01786-2.

[104] Z. T. Crawford, B. Southam, R. Matar, F. R. Avilucea, K. Bowers, M. Altaye, and M. T. Archdeacon, "A Nomogram for Predicting 30-day Mortality in Elderly Patients Undergoing Hemiarthroplasty for Femoral Neck Fractures," *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 11, p. 2 151 459 320 960 087, Jan. 2020, Publisher: SAGE Publications Inc, ISSN: 2151-4593. DOI: 10.1177/2151459320960087.

[105] M. A. Khan, F. S. Hossain, I. Ahmed, N. Muthukumar, and A. Mohsen, "Predictors of early mortality after hip fracture surgery," *International Orthopaedics*, vol. 37, no. 11, pp. 2119–2124, Nov. 2013, ISSN: 1432-5195. DOI: 10.1007/s00264-013-2068-1.

[106] A. Adunsky, M. Arad, N. Koren-Morag, Y. Fleissig, and E. H. Mizrahi, "Increased 1-year mortality rates among elderly hip fracture patients with atrial fibrillation," *Aging Clinical and Experimental Research*, vol. 24, no. 3, pp. 233–238, Jun. 2012, ISSN: 1720-8319. DOI: 10.1007/BF03325251.

[107] G. B. Aharonoff, K. J. Koval, M. L. Skovron, and J. D. Zuckerman, "Hip fractures in the elderly: predictors of one year mortality," *Journal of orthopaedic trauma*, vol. 11, no. 3, pp. 162–165, 1997. DOI: 10.1097/00005131-199704000-00004.

[108] P. Ariza-Vega, M. T. Kristensen, L. Martín-Martín, and J. J. Jiménez-Moleón, "Predictors of long-term mortality in older people with hip fracture," *Archives of physical medicine and rehabilitation*, vol. 96, no. 7, pp. 1215–1221, 2015. DOI: 10.1016/j.apmr.2015.01.023.

[109] G. Bellelli, P. Mazzola, M. Corsi, A. Zambon, G. Corrao, G. Castoldi, G. Zatti, and G. Annoni, "The combined effect of ADL impairment and delay in time from fracture to surgery on 12-month mortality: an observational study in orthogeriatric patients," *Journal of the American Medical Directors Association*, vol. 13, no. 7, 664–e9, 2012. DOI: 10.1016/j.jamda.2012.06.007.

[110] K. B. Bjorkelund, A. Hommel, K.-G. Thorngren, D. Lundberg, and S. Larsson, "Factors at admission associated with 4 months outcome in elderly patients with hip fracture," *AANA Journal-American Association of NurseAnesthetists*, vol. 77, no. 1, p. 49, 2009. [Online]. Available: https://www.researchgate.net/publication/24179708_Factors_at_admission_associated_with_4_

months_outcome_in_elderly_patients_with_hip_fracture (visited on 07/19/2021).

[111] S. L. Bokshan, S. E. Marcaccio, T. D. Blood, and R. A. Hayda, "Factors influencing survival following hip fracture among octogenarians and nonagenarians in the United States," *Injury*, vol. 49, no. 3, pp. 685–690, 2018. DOI: 10.1016/j.injury.2018.02.004.

[112] J. Carow, J. B. Carow, M. Coburn, B.-S. Kim, B. Bücking, C. Bliemel, L. C. Bollheimer, C. J. Werner, J. P. Bach, and M. Knobe, "Mortality and cardiorespiratory complications in trochanteric femoral fractures: a ten year retrospective analysis," *International orthopaedics*, vol. 41, no. 11, pp. 2371–2380, 2017. DOI: 10.1007/s00264-017-3639-3.

[113] I. S. Cenzer, V. Tang, W. J. Boscardin, A. K. Smith, C. Ritchie, M. I. Wallhagen, R. Espaldon, and K. E. Covinsky, "One-year mortality after hip fracture: development and validation of a prognostic index," *Journal of the American Geriatrics Society*, vol. 64, no. 9, pp. 1863–1868, 2016. DOI: 10.1111/jgs.14237.

[114] J. Elliott, T. Beringer, F. Kee, D. Marsh, C. Willis, and M. Stevenson, "Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery," *Journal of clinical epidemiology*, vol. 56, no. 8, pp. 788–795, 2003. DOI: 10.1016/s0895-4356(03)00129-x.

[115] Y. Endo, G. B. Aharonoff, J. D. Zuckerman, K. A. Egol, and K. J. Koval, "Gender differences in patients with hip fracture: a greater risk of morbidity and mortality in men," *Journal of orthopaedic trauma*, vol. 19, no. 1, pp. 29–35, 2005. DOI: 10.1097/00005131-200501000-00006.

[116] L. Flodin, A. Laurin, J. Lökk, T. Cederholm, and M. Hedström, "Increased 1-year survival and discharge to independent living in overweight hip fracture patients," *Acta orthopaedica*, vol. 87, no. 2, pp. 146–151, 2016. DOI: 10.3109/17453674.2015.1125282.

[117] E. C. Folbert, J. H. Hegeman, M. Vermeer, E. M. Regtuijt, D. van der Velde, H. J. ten Duis, and J. P. Slaets, "Improved 1-year mortality in elderly patients with a hip fracture following integrated orthogeriatric treatment," *Osteoporosis International*, vol. 28, no. 1, pp. 269–277, 2017. DOI: 10.1007/s00198-016-3711-7.

[118] M. J. Giummarra, C. L. Ekegren, J. Gong, P. Simpson, P. A. Cameron, E. Edwards, and B. J. Gabbe, "Twelve month mortality rates and independent living in people aged 65 years or older after isolated hip fracture: A prospective registry-based study," *Injury*, vol. 51, no. 2, pp. 420–428, 2020. DOI: 10.1016/j.injury.2019.11.034.

[119] C. Y. Henderson and J. P. Ryan, "Predicting mortality following hip fracture: an analysis of comorbidities and complications," *Irish Journal of Medical Science (1971-)*, vol. 184, no. 3, pp. 667–671, 2015. DOI: 10.1007/s11845-015-1271-z.

[120] C.-A. Ho, C.-Y. Li, K.-S. Hsieh, and H.-F. Chen, "Factors determining the 1-year survival after operated hip fracture: a hospital-based analysis," *Journal of Orthopaedic Science*, vol. 15, no. 1, pp. 30–37, 2010. DOI: 10.1007/s00776-009-1425-9.

[121] P. Huette, O. Abou-Arab, A.-E. Djebara, B. Terrasi, C. Beyls, P.-G. Guinot, E. Havet, H. Dupont, E. Lorne, A. Ntouba, *et al.*, "Risk factors and mortality of

patients undergoing hip fracture surgery: a one-year follow-up study," *Scientific reports*, vol. 10, no. 1, pp. 1–8, 2020. DOI: 10.1038/s41598-020-66614-5.

[122] L.-W. Hung, Y.-T. Hwang, G.-S. Huang, C.-C. Liang, and J. Lin, "The influence of renal dialysis and hip fracture sites on the 10-year mortality of elderly hip fracture patients: a nationwide population-based observational study," *Medicine*, vol. 96, no. 37, 2017. DOI: 10.1097/MD.0000000000007618.

[123] H. X. Jiang, S. R. Majumdar, D. A. Dick, M. Moreau, J. Raso, D. D. Otto, and D. W. C. Johnston, "Development and initial validation of a risk score for predicting in-hospital and 1-year mortality in patients with hip fractures," *Journal of Bone and Mineral Research*, vol. 20, no. 3, pp. 494–500, 2005. DOI: 10.1359/JBMR.041133.

[124] H.-Y. Kang, K.-h. Yang, Y. N. Kim, S.-h. Moon, W.-J. Choi, D. R. Kang, and S. E. Park, "Incidence and mortality of hip fracture among the elderly population in South Korea: a population-based study using the national health insurance claims data," *BMC public health*, vol. 10, no. 1, pp. 1–9, 2010. DOI: 10.1186/1471-2458-10-230.

[125] S.-M. Kim, Y.-W. Moon, S.-J. Lim, B.-K. Yoon, Y.-K. Min, D.-Y. Lee, and Y.-S. Park, "Prediction of survival, second fracture, and functional recovery following the first hip fracture surgery in elderly patients," *Bone*, vol. 50, no. 6, pp. 1343–1350, 2012. DOI: 10.1016/j.bone.2012.02.633.

[126] P. Mazzola, G. Bellelli, V. Broggini, A. Anzuini, M. Corsi, D. Berruti, F. De Filippi, G. Zatti, and G. Annoni, "Postoperative delirium and pre-fracture disability predict 6-month mortality among the oldest old hip fracture patients," *Aging clinical and experimental research*, vol. 27, no. 1, pp. 53–60, 2015. DOI: 10.1007/s40520-014-0242-y.

[127] R. Menéndez-Colino, T. Alarcon, P. Gotor, R. Queipo, R. Ramírez-Martín, A. Otero, and J. I. González-Montalvo, "Baseline and pre-operative 1-year mortality risk factors in a cohort of 509 hip fracture patients consecutively admitted to a co-managed orthogeriatric unit (FONDA Cohort)," *Injury*, vol. 49, no. 3, pp. 656–661, 2018. DOI: 10.1016/j.injury.2018.01.003.

[128] D. Meng, X. Bai, H. Wu, S. Yao, P. Ren, X. Bai, C. Lu, and Z. Song, "Patient and perioperative factors influencing the functional outcomes and mortality in elderly hip fractures," *Journal of Investigative Surgery*, vol. 34, no. 3, pp. 262–269, 2021. DOI: 10.1080/08941939.2019.1625985.

[129] N. Minicuci, S. Maggi, M. Noale, M. Trabucchi, P. Spolaore, and G. Crepaldi, "Predicting mortality in older patients. The VELCA Study," *Aging clinical and experimental research*, vol. 15, no. 4, pp. 328–335, 2003. DOI: 10.1007/BF03324518.

[130] A. H. Myers, E. G. Robinson, M. L. V. Natta, J. D. Michelson, K. Collins, and S. P. Baker, "Hip fractures among the elderly: factors associated with in-hospital mortality," *American Journal of Epidemiology*, vol. 134, no. 10, pp. 1128–1137, 1991. DOI: 10.1093/oxfordjournals.aje.a116016.

[131] M. Nuotio, P. Tuominen, and T. Luukkaala, "Association of nutritional status as measured by the Mini-Nutritional Assessment Short Form with changes in mobility, institutionalization and death after hip fracture," *European Journal of Clinical Nutrition*, vol. 70, no. 3, pp. 393–398, 2016. DOI: 10.1038/ejcn.2015.174.

[132] B. J. O'Daly, J. C. Walsh, J. F. Quinlan, G. A. Falk, R. Stapleton, W. R. Quinlan, and S. K. O'Rourke, "Serum albumin and total lymphocyte count as predictors of outcome in hip fractures," *Clinical nutrition*, vol. 29, no. 1, pp. 89–93, 2010. DOI: 10.1016/j.clnu.2009.07.007.

[133] A. Padron-Monedero, T. López-Cuadrado, I. Galán, E. V. Martinez-Sanchez, P. Martin, and R. Fernandez-Cuenca, "Effect of comorbidities on the association between age and hospital mortality after fall-related hip fracture in elderly patients," *Osteoporosis International*, vol. 28, no. 5, pp. 1559–1568, 2017. DOI: 10.1007/s00198-017-3926-2.

[134] S. R. M. Pereira, M. T. E. Puts, M. C. Portela, and M. A. Sayeg, "The impact of prefracture and hip fracture characteristics on mortality in older persons in Brazil," *Clinical Orthopaedics and Related Research®*, vol. 468, no. 7, pp. 1869–1883, 2010. DOI: 10.1007/s11999-009-1147-5.

[135] M. B. Petersen, H. L. Jørgensen, K. Hansen, and B. R. Duus, "Factors affecting postoperative mortality of patients with displaced femoral neck fracture," *Injury*, vol. 37, no. 8, pp. 705–711, 2006. DOI: 10.1016/j.injury.2006.02.046.

[136] T. A. Ribeiro, M. O. Premaor, J. A. Larangeira, L. G. Brito, M. Luft, L. W. Guterres, and O. A. Monticielo, "Predictors of hip fracture mortality at a general hospital in South Brazil: an unacceptable surgical delay," *Clinics*, vol. 69, pp. 253–258, 2014. DOI: 10.6061/clinics/2014(04)06.

[137] F. Rosso, F. Dettoni, D. E. Bonasia, F. Olivero, L. Mattei, M. Bruzzone, A. Marmotti, and R. Rossi, "Prognostic factors for mortality after hip fracture: operation within 48 hours is mandatory," *Injury*, vol. 47, S91–S97, 2016. DOI: 10.1016/j.injury.2016.07.055.

[138] J. Sanz-Reig, J. S. Marín, J. F. Martínez, D. O. Beltrán, J. F. M. López, and J. A. Q. Rico, "Prognostic factors and predictive model for in-hospital mortality following hip fractures in the elderly," *Chinese Journal of Traumatology*, vol. 21, no. 3, pp. 163–169, 2018. DOI: 10.1016/j.cjtee.2017.10.006.

[139] A. Söderqvist, W. Ekström, S. Ponzer, H. Pettersson, T. Cederholm, N. Dalén, M. Hedström, and J. Tidermark, "Prediction of mortality in elderly patients with hip fractures: a two-year prospective study of 1,944 patients," *Gerontology*, vol. 55, no. 5, pp. 496–504, 2009. DOI: 10.1159/000230587.

[140] S. Tal, A. Gurevich, S. Sagiv, and V. Guller, "Predictors of mortality in hip fracture patients," *European Geriatric Medicine*, vol. 7, no. 6, pp. 561–565, 2016. DOI: 10.1007/BF03324834.

[141] O. Talsnes, F. Hjelmstedt, O. E. Dahl, A. H. Pripp, and O. Reikerås, "Clinical and biochemical prediction of early fatal outcome following hip fracture in the elderly," *International orthopaedics*, vol. 35, no. 6, pp. 903–907, 2011. DOI: 10.1007/s00264-010-1149-7.

[142] A. R. Vosoughi, M. J. Emami, B. Pourabbas, and H. Mahdaviazad, "Factors increasing mortality of the elderly following hip fracture surgery: role of body mass index, age, and smoking," *MUSCULOSKELETAL SURGERY*, vol. 101, no. 1, pp. 25–29, Apr. 2017, ISSN: 2035-5114. DOI: 10.1007/s12306-016-0432-1. [Online]. Available: https://doi.org/10.1007/s12306-016-0432-1 (visited on 03/09/2022).

[143] F. Xing, R. Luo, W. Chen, and X. Zhou, "The risk-adjusted Charlson comorbidity index as a new predictor of one-year mortality rate in elderly Chinese patients who underwent hip fracture surgery," *Orthopaedics & Traumatology: Surgery & Research*, vol. 107, no. 3, p. 102 860, 2021. DOI: 10.1016/j.otsr.2021.102860.

[144] J. C. Yombi, D. C. Putineanu, O. Cornu, P. Lavand'homme, P. Cornette, and D. Castanares-Zapatero, "Low haemoglobin at admission is associated with mortality after hip fractures in elderly patients," *The bone & joint journal*, vol. 101, no. 9, pp. 1122–1128, 2019. DOI: 10.1302/0301-620X.101B9.BJJ-2019-0526.R1.

[145] P. K. Baidoo, J. B. Odei, V. Ansu, M. Segbefia, and H. Holdbrook-Smith, "Predictors of hip fracture mortality in Ghana: a single-center prospective study," *Archives of Osteoporosis*, vol. 16, no. 1, pp. 1–8, 2021. DOI: 10.1007/s11657-021-00883-z.

[146] S. Camur and H. Celik, "Prediction of the Mortality with Comorbidity-Polypharmacy Score in the Osteoporotic Hip Fractures.," *Acta chirurgiae orthopaedicae et traumatologiae Cechoslovaca*, vol. 86, no. 5, pp. 320–323, 2019. [Online]. Available: http://www.achot.cz/dwnld/achot_2019_5_320_323.pdf (visited on 11/10/2021).

[147] Y.-P. Chen, Y.-J. Kuo, C.-h. Liu, P.-C. Chien, W.-C. Chang, C.-Y. Lin, and A. H. Pakpour, "Prognostic factors for 1-year functional outcome, quality of life, care demands, and mortality after surgery in Taiwanese geriatric patients with a hip fracture: a prospective cohort study," *Therapeutic Advances in Musculoskeletal Disease*, vol. 13, pp. 1–11, 2021. DOI: 10.1177/1759720X211028360.

[148] D.-A. Eschbach, L. Oberkircher, C. Bliemel, J. Mohr, S. Ruchholtz, and B. Buecking, "Increased age is not associated with higher incidence of complications, longer stay in acute care hospital and in hospital mortality in geriatric hip fracture patients," *Maturitas*, vol. 74, no. 2, pp. 185–189, 2013. DOI: 10.1016/j.maturitas.2012.11.003.

[149] G. Fu, M. Li, Y. Xue, H. Wang, R. Zhang, Y. Ma, and Q. Zheng, "Rapid preoperative predicting tools for 1-year mortality and walking ability of Asian elderly femoral neck fracture patients who planned for hip arthroplasty," *Journal of Orthopaedic Surgery and Research*, vol. 16, no. 1, pp. 1–11, 2021. DOI: 10.1186/s13018-021-02605-0.

[150] A. A. Mangoni, B. C. van Munster, R. J. Woodman, and S. E. de Rooij, "Measures of anticholinergic drug exposure, serum anticholinergic activity, and all-cause postdischarge mortality in older hospitalized patients with hip fractures," *The American Journal of Geriatric Psychiatry*, vol. 21, no. 8, pp. 785–793, 2013. DOI: 10.1016/j.jagp.2013.01.012.

[151] A. Söderqvist, R. Miedel, S. Ponzer, and J. Tidermark, "The influence of cognitive function on outcome after a hip fracture," *The Journal of Bone and Joint Surgery. American Volume*, vol. 88, no. 10, pp. 2115–2123, Oct. 2006, ISSN: 0021-9355. DOI: 10.2106/JBJS.E.01409.

[152] J.-I. Yoo, H. Kim, Y.-C. Ha, H.-B. Kwon, and K.-H. Koo, "Osteosarcopenia in patients with hip fracture is related with high mortality," *Journal of Korean medical science*, vol. 33, no. 4, 2018. DOI: 10.3346/jkms.2018.33.e27.

[153] J. J. Bell, R. C. Pulle, A. M. Crouch, S. S. Kuys, R. L. Ferrier, and S. L. Whitehouse, "Impact of malnutrition on 12-month mortality following acute hip fracture," *ANZ journal of surgery*, vol. 86, no. 3, pp. 157–161, 2016. DOI: 10.1111/ans.13429.

[154] C. Bliemel, R. Sielski, B. Doering, R. Dodel, M. Balzer-Geldsetzer, S. Ruchholtz, and B. Buecking, "Pre-fracture quality of life predicts 1-year survival in elderly patients with hip fracture—development of a new scoring system," *Osteoporosis international*, vol. 27, no. 6, pp. 1979–1987, 2016. DOI: 10.1007/s00198-015-3472-8.

[155] Y. Ishidou, C. Koriyama, H. Kakoi, T. Setoguchi, S. Nagano, M. Hirotsu, T. Yamamoto, M. Yokouchi, and S. Komiya, "Predictive factors of mortality and deterioration in performance of activities of daily living after hip fracture surgery in Kagoshima, Japan," *Geriatrics & gerontology international*, vol. 17, no. 3, pp. 391–401, 2017. DOI: 10.1111/ggi.12718.

[156] P. N. Kannegaard, S. van der Mark, P. Eiken, and B. O. Abrahamsen, "Excess mortality in men compared with women following a hip fracture. National analysis of comedications, comorbidity and survival," *Age and ageing*, vol. 39, no. 2, pp. 203–209, 2010. DOI: 10.1093/ageing/afp221.

[157] G. J. Heyes, A. Tucker, D. Marley, and A. Foster, "Predictors for 1-year mortality following hip fracture: a retrospective review of 465 consecutive patients," *European journal of trauma and emergency surgery*, vol. 43, no. 1, pp. 113–119, 2017. DOI: 10.1007/s00068-015-0556-2.

[158] A. Lizaur-Utrilla, D. Martinez-Mendez, I. Collados-Maestre, F. A. Miralles-Muñoz, L. Marco-Gomez, and F. A. Lopez-Prats, "Early surgery within 2 days for hip fracture is not reliable as healthcare quality indicator," *Injury*, vol. 47, no. 7, pp. 1530–1535, 2016. DOI: 10.1016/j.injury.2016.04.040.

[159] G. Pioli, A. Barone, A. Giusti, M. Oliveri, M. Pizzonia, M. Razzano, and E. Palummeri, "Predictors of mortality after hip fracture: results from 1-year follow-up," *Aging clinical and experimental research*, vol. 18, no. 5, pp. 381–387, 2006. DOI: 10.1007/BF03324834.

[160] B.-G. Kim, Y.-K. Lee, H.-P. Park, H.-M. Sohn, A.-Y. Oh, Y.-T. Jeon, and K.-H. Koo, "C-reactive protein is an independent predictor for 1-year mortality in elderly patients undergoing hip fracture surgery: A retrospective analysis," *Medicine*, vol. 95, no. 43, 2016. DOI: 10.1097/MD.0000000000005152.

[161] M. Mariconda, G. G. Costa, S. Cerbasi, P. Recano, E. Aitanti, M. Gambacorta, and M. Misasi, "The determinants of mortality and morbidity during the year following fracture of the hip: a prospective study," *The bone & joint journal*, vol. 97, no. 3, pp. 383–390, 2015. DOI: 10.1302/0301-620X.97B3.34504.

[162] Y. Camurcu, A. Cobden, H. Sofu, N. Saklavci, and M. Kis, "What are the determinants of mortality after cemented bipolar hemiarthroplasty for unstable intertrochanteric fractures in elderly patients?" *The Journal of arthroplasty*, vol. 32, no. 10, pp. 3038–3043, 2017. DOI: 10.1016/j.arth.2017.04.042.

[163] G. Thorne and L. Hodgson, "Performance of the Nottingham Hip Fracture Score and Clinical Frailty Scale as predictors of short and long-term outcomes: a dual-centre 3-year observational study of hip fracture patients," *Journal of Bone and Mineral Metabolism*, vol. 39, no. 3, pp. 494–500, 2021. DOI: 10.1007/s00774-020-01187-x.

[164] M. Härstedt, C. Rogmark, R. Sutton, O. Melander, and A. Fedorowski, "Impact of comorbidity on 6-month hospital readmission and mortality after hip fracture surgery," *Injury*, vol. 46, no. 4, pp. 713–718, 2015. DOI: 10.1016/j.injury.2014.12.024.

[165] M. Zanetti, G. G. Cappellari, C. Ratti, G. Ceschia, L. Murena, P. De Colle, and R. Barazzoni, "Poor nutritional status but not cognitive or functional impairment per se independently predict 1 year mortality in elderly patients with hip-fracture," *Clinical Nutrition*, vol. 38, no. 4, pp. 1607–1612, 2019. DOI: 10.1016/j.clnu.2018.08.030.

[166] T. W. Lau, C. X. Fang, and F. K. L. Leung, "Assessment of postoperative short-term and long-term mortality risk in Chinese geriatric patients for hip fracture using the Charlson comorbidity score," *Hong Kong medical journal*, 2015. DOI: 10.12809/hkmj154451.

[167] L.-C. Wu, M.-Y. Chou, C.-K. Liang, Y.-T. Lin, Y.-C. Ku, and R.-H. Wang, "Factors affecting one-year mortality of elderly patients after surgery for hip fracture," *International Journal of Gerontology*, vol. 10, no. 4, pp. 207–211, 2016. DOI: 10.1016/j.ijge.2016.02.004.

[168] F. D'Angelo, M. Giudici, M. Molina, and G. Margaria, "Mortality rate after hip hemiarthroplasty: analysis of risk factors in 299 consecutives cases," *Journal of Orthopaedics and Traumatology*, vol. 6, no. 3, pp. 111–116, Oct. 2005, ISSN: 1590-9921, 1590-9999. DOI: 10.1007/s10195-005-0093-6. [Online]. Available: http://link.springer.com/10.1007/s10195-005-0093-6 (visited on 03/08/2022).

[169] A. Fisher, L. Fisher, W. Srikusalanukul, and P. N. Smith, "Usefulness of simple biomarkers at admission as independent indicators and predictors of in-hospital mortality in older hip fracture patients," *Injury*, vol. 49, no. 4, pp. 829–840, 2018. DOI: 10.1016/j.injury.2018.03.005.

[170] M. Velez, U. Palacios-Barahona, M. Paredes-Laverde, and J. A. Ramos-Castaneda, "Factors associated with mortality due to trochanteric fracture. A cross-sectional study," *Orthopaedics & Traumatology: Surgery & Research*, vol. 106, no. 1, pp. 135–139, 2020. DOI: 10.1016/j.otsr.2019.06.022.

[171] F. M. Kovar, G. Endler, O. F. Wagner, and M. Jaindl, "Basal haemoglobin levels as prognostic factor for early death in elderly patients with a hip fracture–a twenty year observation study," *Injury*, vol. 46, no. 6, pp. 1018–1022, 2015. DOI: 10.1016/j.injury.2015.01.010.

[172] S. Aldebeyan, A. Nooh, A. Aoude, M. H. Weber, and E. J. Harvey, "Hypoalbuminaemia a marker of malnutrition and predictor of postoperative complications and mortality after hip fractures," *Injury*, vol. 48, no. 2, pp. 436–440, 2017. DOI: 10.1016/j.injury.2016.12.016.

[173] D. Norring-Agerskov, C. M. Madsen, B. Abrahamsen, T. Riis, O. B. Pedersen, N. R. Jørgensen, L. Bathum, J. B. Lauritzen, and H. L. Jørgensen, "Hyperkalemia is associated with increased 30-day mortality in hip fracture patients," *Calcified tissue international*, vol. 101, no. 1, pp. 9–16, 2017. DOI: 10.1007/s00223-017-0252-9.

[174] J. C. Ho, "Interruptions: using activity transitions to trigger proactive messages," M.S. thesis, Massachusetts Institute of Technology, 2004. [Online]. Available: http://hdl.handle.net/1721.1/33135 (visited on 08/10/2021).

[175] J. E. Lawrence, D. M. Fountain, D. J. Cundall-Curry, and A. D. Carrothers, "Do patients taking warfarin experience delays to theatre, longer hospital stay, and poorer survival after hip fracture?" *Clinical Orthopaedics and Related Research®*, vol. 475, no. 1, pp. 273–279, 2017. DOI: 10.1007/s11999-016-5056-0.

[176] I. Moppett, M. Parker, R. Griffiths, T. Bowers, S. White, and C. Moran, "Nottingham Hip Fracture Score: longitudinal and multi-centre assessment," *British Journal of Anaesthesia*, vol. 109, no. 4, pp. 546–550, Oct. 2012, ISSN: 00070912. DOI: 10.1093/bja/aes187.

[177] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, Jan. 1987, ISSN: 0021-9681. DOI: 10.1016/0021-9681(87)90171-8.

[178] L. A. Stevens, G. Viswanathan, and D. E. Weiner, "Chronic kidney disease and end-stage renal disease in the elderly population: current prevalence, future projections, and clinical significance," *Advances in Chronic Kidney Disease*, vol. 17, no. 4, pp. 293–301, Jul. 2010, ISSN: 1548-5609. DOI: 10.1053/j.ackd.2010.03.010.

[179] N. K. Sweitzer, M. Lopatin, C. W. Yancy, R. M. Mills, and L. W. Stevenson, "Comparison of clinical features and outcomes of patients hospitalized with heart failure and normal ejection fraction ($>$ or $=$55%) versus those with mildly reduced (40% to 55%) and moderately to severely reduced ($<$40%) fractions," *The American Journal of Cardiology*, vol. 101, no. 8, pp. 1151–1156, Apr. 2008, ISSN: 0002-9149. DOI: 10.1016/j.amjcard.2007.12.014.

[180] M. Simmonds, "Quantifying the risk of error when interpreting funnel plots," *Systematic reviews*, vol. 4, no. 1, pp. 1–7, 2015. DOI: 10.1186/s13643-015-0004-8.

[181] T. Ogawa, H. Schermann, H. Kobayashi, K. Fushimi, A. Okawa, and T. Jinno, "Age and clinical outcomes after hip fracture surgery: do octogenarian, nonagenarian and centenarian classifications matter?" *Age and Ageing*, vol. 50, no. 6, pp. 1952–1960, Nov. 2021, ISSN: 0002-0729. DOI: 10.1093/ageing/afab137.

[182] S.-J. Kim, H.-S. Park, and D.-W. Lee, "Outcome of nonoperative treatment for hip fractures in elderly patients: A systematic review of recent literature," *Journal of Orthopaedic Surgery*, vol. 28, no. 2, p. 2 309 499 020 936 848, May 2020, Publisher: SAGE Publications Ltd STM, ISSN: 1022-5536. DOI: 10.1177/2309499020936848.

[183] S. D. Berry, R. R. Rothbaum, D. P. Kiel, Y. Lee, and S. L. Mitchell, "Association of Clinical Outcomes With Surgical Repair of Hip Fracture vs Nonsurgical Management in Nursing Home Residents With Advanced Dementia," *JAMA Internal Medicine*, vol. 178, no. 6, pp. 774–780, Jun. 2018, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2018.0743.

[184] J. T. Clapp, M. L. Schwarze, and L. A. Fleisher, "Surgical Overtreatment and Shared Decision-making—The Limits of Choice," *JAMA Surgery*, vol. 157, no. 1, pp. 5–6, Jan. 2022, ISSN: 2168-6254. DOI: 10.1001/jamasurg.2021.4425.

[185] M. Cardona-Morrell, J. Kim, R. M. Turner, M. Anstey, I. A. Mitchell, and K. Hillman, "Non-beneficial treatments in hospital at the end of life: a systematic review on extent of the problem," *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, vol. 28, no. 4, pp. 456–469, Sep. 2016, ISSN: 1464-3677. DOI: 10.1093/intqhc/mzw060.

[186] H. H. Wijnen, P. P. Schmitz, H. Es-Safraouy, L. A. Roovers, D. G. Taekema, and J. L. C. Van Susante, "Nonoperative management of hip fractures in very frail elderly patients may lead to a predictable short survival as part of advance care planning," *Acta Orthopaedica*, vol. 92, no. 6, pp. 728–732, Dec. 2021, ISSN: 1745-3682. DOI: 10.1080/17453674.2021.1959155.

[187] F. C. Ko and R. S. Morrison, "Hip Fracture: A Trigger for Palliative Care in Vulnerable Older Adults," *JAMA Internal Medicine*, vol. 174, no. 8, pp. 1281–1282, Aug. 2014, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2014.999.

[188] M. D. Neuman, J. H. Silber, J. S. Magaziner, M. A. Passarella, S. Mehta, and R. M. Werner, "Survival and Functional Outcomes After Hip Fracture Among Nursing Home Residents," *JAMA Internal Medicine*, vol. 174, no. 8, pp. 1273–1280, Aug. 2014, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2014.2362.

[189] B. C. van der Zwaard, C. E. Stein, J. E. M. Bootsma, H. J. A. A. van Geffen, C. M. Douw, and C. J. P. W. Keijsers, "Fewer patients undergo surgery when adding a comprehensive geriatric assessment in older patients with a hip fracture," *Archives of Orthopaedic and Trauma Surgery*, vol. 140, no. 4, pp. 487–492, Apr. 2020, ISSN: 1434-3916. DOI: 10.1007/s00402-019-03294-5.

[190] P. Prommik, K. Tootsi, T. Saluse, A. Märtson, and H. Kolk, "Nonoperative hip fracture management practices and patient survival compared to surgical care: an analysis of Estonian population-wide data," *Archives of Osteoporosis*, vol. 16, no. 1, p. 101, Dec. 2021, ISSN: 1862-3522, 1862-3514. DOI: 10.1007/s11657-021-00973-y.

[191] B. Yenidogan, S. Pathak, J. Geerdink, J. H. Hegeman, and M. van Keulen, "Multimodal Machine Learning for 30-Days Post-Operative Mortality Prediction of Elderly Hip Fracture Patients," in *2021 International Conference on Data Mining Workshops (ICDMW)*, ISSN: 2375-9259, Dec. 2021, pp. 508–516. DOI: 10.1109/ICDMW53433.2021.00068.

[192] M. P. Cary, F. Zhuang, R. L. Draelos, W. Pan, S. Amarasekara, B. J. Douthit, Y. Kang, and C. S. Colón-Emeric, "Machine Learning Algorithms to Predict Mortality and Allocate Palliative Care for Older Patients With Hip Fracture," *Journal of the American Medical Directors Association*, vol. 22, no. 2, pp. 291–296, Feb. 2021, ISSN: 1525-8610. DOI: 10.1016/j.jamda.2020.09.025.

[193] S. French, A. M. Hanea, T. Bedford, and G. F. Nane, "Introduction and Overview of Structured Expert Judgement," in *Expert Judgement in Risk and Decision Analysis*, A. M. Hanea, G. F. Nane, T. Bedford, and S. French, Eds., ser. International Series in Operations Research & Management Science. Cham: Springer, 2021, vol. 293, pp. 1–16, ISBN: 978-3-030-46473-8. DOI: 10.1007/978-3-030-46474-5_1.

[194] A. M. Hanea and G. F. Nane, "Calibrating experts' probabilistic assessments for improved probabilistic predictions," *Safety Science*, vol. 118, pp. 763–771, Oct. 2019, ISSN: 0925-7535. DOI: 10.1016/j.ssci.2019.05.048.

[195] R. H. Quinn, P. A. Mooar, J. N. Murray, R. Pezold, and K. S. Sevarino, "Treatment of Hip Fractures in the Elderly," *The Journal of the American Academy of Orthopaedic Surgeons*, vol. 25, no. 5, e102–e104, May 2017, ISSN: 1940-5480. DOI: 10.5435/JAAOS-D-16-00431.

[196] M. Ryan and S. Farrar, "Using conjoint analysis to elicit preferences for health care," *BMJ*, vol. 320, no. 7248, pp. 1530–1533, Jun. 2000, Publisher: British Medical Journal Publishing Group Section: Education and debate, ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.320.7248.1530.

[197] M. Ryan, D. A. Scott, C. Reeves, A. Bate, E. R. van Teijlingen, E. M. Russell, M. Napper, and C. M. Robb, "Eliciting public preferences for healthcare: a systematic review of techniques," *Health Technology Assessment*, vol. 5, no. 5, Mar. 2001, ISSN: ISSN: 2046-4924, ISSN: 1366-5278. DOI: 10.3310/hta5050.

[198] L. M. Bachmann, A. Mühleisen, A. Bock, G. ter Riet, U. Held, and A. G. H. Kessels, "Vignette studies of medical choice and judgement to study caregivers' medical decision behaviour: systematic review," *BMC medical research methodology*, vol. 8, no. 1, pp. 1–8, 2008. DOI: 10.1186/1471-2288-8-50.

[199] C. Atzmüller and P. M. Steiner, "Experimental vignette studies in survey research," *Methodology*, 2010. DOI: 10.1027/1614-2241/a000014.

[200] S. C. Evans, M. C. Roberts, J. W. Keeley, J. B. Blossom, C. M. Amaro, A. M. Garcia, C. O. Stough, K. S. Canter, R. Robles, and G. M. Reed, "Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies," *International journal of clinical and health psychology*, vol. 15, no. 2, pp. 160–170, 2015. DOI: 10.1016/j.ijchp.2014.12.001.

[201] B. J. Taylor, "Factorial Surveys: Using Vignettes to Study Professional Judgement," *The British Journal of Social Work*, vol. 36, no. 7, pp. 1187–1207, Oct. 2006, ISSN: 0045-3102. DOI: 10.1093/bjsw/bch345.

[202] R. M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, 1991, ISBN: 0-19-506465-8.

[203] R. de Groot, W. S. Nijmeijer, E. C. Folbert, M. M. R. Vollenbroek-Hutten, and J. H. Hegeman, "'Nonagenarians' with a hip fracture: is a different orthogeriatric treatment strategy necessary?" *Archives of osteoporosis*, vol. 15, no. 1, pp. 1–9, 2020. DOI: 10.1007/s11657-020-0698-7.

[204] E. E. Hurwitz, M. Simon, S. R. Vinta, C. F. Zehm, S. M. Shabot, A. Minhajuddin, and A. E. Abouleish, "Adding Examples to the ASA-Physical Status Classification Improves Correct Assignment to Patients," *Anesthesiology*, vol. 126, no. 4, pp. 614–622, Apr. 2017, ISSN: 0003-3022. DOI: 10.1097/ALN.0000000000001541.

[205] D. Mayhew, V. Mendonca, and B. V. S. Murthy, "A review of ASA physical status – historical perspectives and modern developments," *Anaesthesia*, vol. 74, no. 3, pp. 373–379, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/anae.14569, ISSN: 1365-2044. DOI: 10.1111/anae.14569.

[206] U. Lucca, M. Garrì, A. Recchia, G. Logroscino, P. Tiraboschi, M. Franceschi, C. Bertinotti, A. Biotti, E. Gargantini, M. Maragna, A. Nobili, L. Pasina, C. Franchi, E. Riva, and M. Tettamanti, "A Population-based study of dementia in the oldest old: the Monzino 80-plus Study," *BMC Neurology*, vol. 11, no. 1, p. 54, Dec. 2011, ISSN: 1471-2377. DOI: 10.1186/1471-2377-11-54.

[207] D. Mungas, B. R. Reed, W. G. Ellis, and W. J. Jagust, "The Effects of Age on Rate of Progression of Alzheimer Disease and Dementia With Associated Cerebrovascular Disease," *Archives of Neurology*, vol. 58, no. 8, pp. 1243–1247, Aug. 2001, ISSN: 0003-9942. DOI: 10.1001/archneur.58.8.1243.

[208] B. L. d. C. Araujo and D. Theobald, "Letter to the Editor: ASA Physical Status Classification in Surgical Oncology and the Importance of Improving Inter-Rater Reliability," *Journal of Korean Medical Science*, vol. 32, no. 7, pp. 1211–1212, Jul. 2017, ISSN: 1011-8934. DOI: 10.3346/jkms.2017.32.7.1211.

[209] F. R. Johnson, E. Lancsar, D. Marshall, V. Kilambi, A. Mühlbacher, D. A. Regier, B. W. Bresnahan, B. Kanninen, and J. F. Bridges, "Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force," *Value in Health*, vol. 16, no. 1, pp. 3–13, 2013, ISSN: 1098-3015. DOI: https://doi.org/10.1016/j.jval.2012.08.2223. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1098301512041629.

[210] P. de Aguiar, B. Bourguignon, M. Khots, D. Massart, and R. Phan-Than-Luu, "D-optimal designs," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 2, pp. 199–210, 1995, ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(94)00076-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016974399400076X.

[211] T. Morgan-Wall and G. Khoury, "Optimal Design Generation and Power Evaluation in R: The skpr Package," *Journal of Statistical Software*, vol. 99, no. 1, pp. 1–36, 2021. DOI: 10.18637/jss.v099.i01.

[212] A. B. Hauber, J. M. González, C. G. M. Groothuis-Oudshoorn, T. Prior, D. A. Marshall, C. Cunningham, M. J. IJzerman, and J. F. P. Bridges, "Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force," *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, vol. 19, no. 4, pp. 300–315, Jun. 2016, ISSN: 1524-4733. DOI: 10.1016/j.jval.2016.04.004.

[213] A. Ali, S. Ali, S. A. Khan, D. M. Khan, K. Abbas, A. Khalil, S. Manzoor, and U. Khalil, "Sample size issues in multilevel logistic regression models," *PLoS ONE*, vol. 14, no. 11, e0225427, Nov. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0225427.

[214] D. McNeish, "Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction," *Multivariate Behavioral Research*, vol. 52, no. 5, pp. 661–670, Sep. 2017, ISSN: 0027-3171. DOI: 10.1080/00273171.2017.1344538.

[215] ——, "On using Bayesian methods to address small sample problems," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 23, no. 5, pp. 750–773, 2016. DOI: 10.1080/10705511.2016.1186549.

[216] G. Hamra, R. MacLehose, and D. Richardson, "Markov Chain Monte Carlo: an introduction for epidemiologists," *International Journal of Epidemiology*, vol. 42, no. 2, pp. 627–634, Apr. 2013, ISSN: 1464-3685, 0300-5771. DOI: 10.1093/ije/dyt043. (visited on 08/20/2022).

[217] S. C. Smid, D. McNeish, M. Miočević, and R. van de Schoot, "Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 27, no. 1, pp. 131–161, Jan. 2020, ISSN: 1070-5511. DOI: 10.1080/10705511.2019.1577140.

[218] D. A. Harville, "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–338, Jun. 1977, ISSN: 0162-1459. DOI: 10.1080/01621459.1977.10480998.

[219] M. G. Kenward and J. H. Roger, "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, vol. 53, no. 3, pp. 983–997, 1997, Publisher: [Wiley, International Biometric Society], ISSN: 0006-341X. DOI: 10.2307/2533558.

[220] D. A. Regier, M. Ryan, E. Phimister, and C. A. Marra, "Bayesian and classical estimation of mixed logit: An application to genetic testing," *Journal of Health Economics*, vol. 28, no. 3, pp. 598–610, May 2009, ISSN: 01676296. DOI: 10.1016/j.jhealeco.2008.11.003.

[221] M. Miočević, R. Levy, and R. v. d. Schoot, "Introduction to Bayesian Statistics," in *Small Sample Size Solutions*, Routledge, 2020, ISBN: 978-0-429-27387-2.

[222] S. Depaoli and R. van de Schoot, "Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist," *Psychological Methods*, vol. 22, no. 2, pp. 240–261, 2017, ISSN: 1939-1463. DOI: 10.1037/met0000065.

[223] S. Khan, "Meta-analysis of Odds Ratio," in *Meta-Analysis: Methods for Health and Experimental Studies*, ser. Statistics for Biology and Health, S. Khan, Ed., Singapore: Springer, 2020, pp. 87–118, ISBN: 9789811550324. DOI: 10.1007/978-981-15-5032-4_5.

[224] D. Akinc and M. Vandebroek, "Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix," *Journal of Choice Modelling*, vol. 29, pp. 133–151, Dec. 2018, ISSN: 1755-5345. DOI: 10.1016/j.jocm.2017.11.004.

[225] N. K. Schuurman, R. P. P. P. Grasman, and E. L. Hamaker, "A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models," *Multivariate Behavioral Research*, vol. 51, no. 2-3, pp. 185–206, May 2016, ISSN: 0027-3171, 1532-7906. DOI: 10.1080/00273171.2015.1065398.

[226] Z. Zhang, "A Note on Wishart and Inverse Wishart Priors for Covariance Matrix," *Journal of Behavioral Data Science*, vol. 1, no. 2, 2021, ISSN: 25758306, 25741284. DOI: 10.35566/jbds/v1n2/p2.

[227] J. Lin, M. F. Myers, L. M. Koehly, and C. S. Marcum, "A Bayesian hierarchical logistic regression model of multiple informant family health histories," *BMC Medical Research Methodology*, vol. 19, no. 1, p. 56, Mar. 2019, ISSN: 1471-2288. DOI: 10.1186/s12874-019-0700-5.

[228] W. R. Gilks, "Markov Chain Monte Carlo," in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Eds., John Wiley & Sons, Ltd, 2005, ISBN: 978-0-470-01181-2. DOI: 10.1002/0470011815.b2a14021.

[229] S. Chib and B. P. Carlin, "On MCMC sampling in hierarchical longitudinal models," *Statistics and Computing*, vol. 9, no. 1, pp. 17–26, Apr. 1999, ISSN: 1573-1375. DOI: 10.1023/A:1008853808677.

[230] M. Hein, N. Goeken, P. Kurz, and W. J. Steiner, "Using Hierarchical Bayes draws for improving shares of choice predictions in conjoint simulations: A study based on conjoint choice data," *European Journal of Operational Research*, vol. 297, no. 2, pp. 630–651, Mar. 2022, ISSN: 03772217. DOI: 10.1016/j.ejor.2021.05.056.

[231] A. D. Martin, K. M. Quinn, and J. H. Park, "MCMCpack: Markov Chain Monte Carlo in R," *Journal of Statistical Software*, vol. 42, pp. 1–21, Jun. 2011, ISSN: 1548-7660. DOI: 10.18637/jss.v042.i09.

[232] N. Malhotra, J. Hall, M. Shaw, and P. Oppenheim, *Marketing research : an applied orientation*. Pearson Education Australia, Jan. 2006, ISBN: 978-0-7339-7004-7.

[233] W. S. Nijmeijer, E. C. Folbert, M. Vermeer, M. M. R. Vollenbroek-Hutten, and J. H. Hegeman, "The consistency of care for older patients with a hip fracture: are the results of the integrated orthogeriatric treatment model of the Centre of Geriatric Traumatology consistent 10 years after implementation?" *Archives of Osteoporosis*, vol. 13, no. 1, p. 131, Dec. 2018, ISSN: 1862-3522, 1862-3514. DOI: 10.1007/s11657-018-0550-5.

[234] M. Daabiss, "American Society of Anaesthesiologists physical status classification," *Indian Journal of Anaesthesia*, vol. 55, no. 2, pp. 111–115, 2011, ISSN: 0019-5049. DOI: 10.4103/0019-5049.79879.

[235] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments," *Bayesian statistics*, vol. 4, pp. 641–649, 1992.

[236] E. W. de Bekker-Grob, B. Donkers, M. F. Jonker, and E. A. Stolk, "Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide," *The Patient - Patient-Centered Outcomes Research*, vol. 8, no. 5, pp. 373–384, Oct. 2015, ISSN: 1178-1653, 1178-1661. DOI: 10.1007/s40271-015-0118-z.

[237] R. L. Harrison, "Introduction to Monte Carlo Simulation," *AIP Conference Proceedings*, vol. 1204, no. 1, pp. 17–21, Jan. 2010, Publisher: American Institute of Physics, ISSN: 0094-243X. DOI: 10.1063/1.3295638.

[238] J. F. Bridges, A. B. Hauber, D. Marshall, A. Lloyd, L. A. Prosser, D. A. Regier, F. R. Johnson, and J. Mauskopf, "Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force," *Value in Health*, vol. 14, no. 4, pp. 403–413, 2011, ISSN: 1098-3015. DOI: https://doi.org/10.1016/j.jval.2010.11.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1098301510000835.

[239] P. M. Steiner, C. Atzmüller, and D. Su, "Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap," *Journal of Methods and Measurement in the Social Sciences*, vol. 7, no. 2, Jun. 2017, Number: 2 Publisher: University of Arizona Libraries, ISSN: 2159-7855. DOI: 10.2458/v7i2.20321. (visited on 03/09/2022).

[240] D. Hartley and S. French, "Bayesian Modelling of Dependence Between Experts: Some Comparisons with Cooke's Classical Model," in *Expert Judgement in Risk and Decision Analysis*, A. M. Hanea, G. F. Nane, T. Bedford, and S. French, Eds., ser. International Series in Operations Research & Management Science.

Cham: Springer, Feb. 2021, vol. 293, pp. 115–146, ISBN: 978-3-030-46473-8. DOI: 10.1007/978-3-030-46474-5_5.

[241] T. Hald, W. Aspinall, B. Devleesschauwer, R. Cooke, T. Corrigan, A. H. Havelaar, H. J. Gibb, P. R. Torgerson, M. D. Kirk, F. J. Angulo, R. J. Lake, N. Speybroeck, and S. Hoffmann, "World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation," *PLOS ONE*, vol. 11, no. 1, S. Sreevatsan, Ed., e0145839, Jan. 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0145839.

[242] H. Marti, T. Mazzuchi, and R. Cooke, "Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis," in *Expert Judgement in Risk and Decision Analysis*, A. M. Hanea, G. F. Nane, T. Bedford, and S. French, Eds., ser. International Series in Operations Research & Management Science. Cham: Springer, Feb. 2021, vol. 293, pp. 53–82, ISBN: 978-3-030-46473-8. DOI: 10.1007/978-3-030-46474-5_3.

[243] W. Aspinall, "Structured elicitation of expert judgment for probabilistic hazard and risk assessment in volcanic eruptions," in *Statistics in Volcanology (Special Publications of IAVCEI, No. 1)*, H. Mader, S. Coles, C. Connor, and L. Connor, Eds., Geological Society of London, 2006, pp. 15–30, ISBN: 978-1-86239-208-3.

[244] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, pp. 1–67, Dec. 2011, ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03.

[245] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Jun. 2004, ISBN: 978-0-471-65574-9.

[246] A. M. Hanea and G. F. Nane, "An In-Depth Perspective on the Classical Model," in *Expert Judgement in Risk and Decision Analysis*, ser. International Series in Operations Research & Management Science, A. M. Hanea, G. F. Nane, T. Bedford, and S. French, Eds., Cham: Springer International Publishing, 2021, pp. 225–256, ISBN: 978-3-030-46474-5. DOI: 10.1007/978-3-030-46474-5_10.

[247] N. Sommet and D. Morselli, "Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS," *International Review of Social Psychology*, vol. 30, no. 1, pp. 203–218, Sep. 2017, ISSN: 2397-8570. DOI: 10.5334/irsp.90.

[248] G. A. Sachs, J. W. Shega, and D. Cox-Hayley, "Barriers to Excellent End-of-life Care for Patients with Dementia," *Journal of General Internal Medicine*, vol. 19, no. 10, pp. 1057–1063, Oct. 2004, ISSN: 0884-8734. DOI: 10.1111/j.1525-1497.2004.30329.x.

[249] S. L. Mitchell, D. K. Kiely, and M. B. Hamel, "Dying with advanced dementia in the nursing home," *Archives of Internal Medicine*, vol. 164, no. 3, pp. 321–326, Feb. 2004, ISSN: 0003-9926. DOI: 10.1001/archinte.164.3.321.

[250] J. C. Ahronheim, R. S. Morrison, S. A. Baskin, J. Morris, and D. E. Meier, "Treatment of the dying in the acute care hospital. Advanced dementia and metastatic cancer," *Archives of Internal Medicine*, vol. 156, no. 18, pp. 2094–2100, Oct. 1996, ISSN: 0003-9926. DOI: 10.1001/archinte.1996.00440170110012.

[251] K. Boyd and S. A. Murray, "Recognising and managing key transitions in end of life care," *BMJ*, vol. 341, p. c4863, Sep. 2010, ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.c4863.

[252] M. L. Rurup, B. D. Onwuteaka-Philipsen, H. R. W. Pasman, M. W. Ribbe, and G. van der Wal, "Attitudes of physicians, nurses and relatives towards end-of-life decisions concerning nursing home patients with dementia," *Patient Education and Counseling*, vol. 61, no. 3, pp. 372–380, Jun. 2006, ISSN: 0738-3991. DOI: 10.1016/j.pec.2005.04.016.

[253] L. Volicer, Y. Rheaume, J. Brown, K. Fabiszewski, and R. Brady, "Hospice Approach to the Treatment of Patients With Advanced Dementia of the Alzheimer Type," *JAMA*, vol. 256, no. 16, pp. 2210–2213, Oct. 1986, ISSN: 0098-7484. DOI: 10.1001/jama.1986.03380160068022.

[254] T. O. Smith, A. Cooper, G. Peryer, R. Griffiths, C. Fox, and J. Cross, "Factors predicting incidence of post-operative delirium in older people following hip fracture surgery: a systematic review and meta-analysis," *International Journal of Geriatric Psychiatry*, vol. 32, no. 4, pp. 386–396, 2017, ISSN: 1099-1166. DOI: 10.1002/gps.4655.

[255] M. Bitsch, N. Foss, B. Kristensen, and H. Kehlet, "Pathogenesis of and management strategies for postoperative delirium after hip fracture A review," *Acta Orthopaedica Scandinavica*, vol. 75, no. 4, pp. 378–389, Jan. 2004, ISSN: 0001-6470. DOI: 10.1080/00016470410001123.

[256] C. A. Mosk, M. Mus, J. P. A. M. Vroemen, T. van der Ploeg, D. I. Vos, L. H. G. J. Elmans, and L. van der Laan, "Dementia and delirium, the outcomes in elderly hip fracture patients," *Clinical Interventions in Aging*, vol. 12, pp. 421–430, Mar. 2017, ISSN: 1176-9092. DOI: 10.2147/CIA.S115945.

[257] S. B. Yellen, D. F. Cella, and W. T. Leslie, "Age and Clinical Decision Making in Oncology Patients," *JNCI: Journal of the National Cancer Institute*, vol. 86, no. 23, pp. 1766–1770, Dec. 1994, ISSN: 0027-8874. DOI: 10.1093/jnci/86.23.1766.

[258] E. Bastiaannet, G. J. Liefers, A. J. M. de Craen, P. J. K. Kuppen, W. van de Water, J. E. A. Portielje, L. G. M. van der Geest, M. L. G. Janssen-Heijnen, O. M. Dekkers, C. J. H. van de Velde, and R. G. J. Westendorp, "Breast cancer in elderly compared to younger patients in the Netherlands: stage at diagnosis, treatment and survival in 127,805 unselected patients," *Breast Cancer Research and Treatment*, vol. 124, no. 3, pp. 801–807, Dec. 2010, ISSN: 0167-6806, 1573-7217. DOI: 10.1007/s10549-010-0898-8.

[259] K. Adamowicz and Z. Baczkowska-Waliszewska, "Quality of life during chemotherapy, hormonotherapy or antiHER2 therapy of patients with advanced, metastatic breast cancer in clinical practice," *Health and Quality of Life Outcomes*, vol. 18, no. 1, p. 134, May 2020, ISSN: 1477-7525. DOI: 10.1186/s12955-020-01389-x.

[260] C. Falci, E. Morello, and J. P. Droz, "Treatment of prostate cancer in unfit senior adult patients," *Cancer Treatment Reviews*, Therapeutic Approach for Unfit Older Patients in the Principal Tumor Types, vol. 35, no. 6, pp. 522–527, Oct. 2009, ISSN: 0305-7372. DOI: 10.1016/j.ctrv.2009.04.014.

[261] D. Hind, L. Wyld, and M. W. Reed, "Surgery, with or without tamoxifen, vs tamoxifen alone for older women with operable breast cancer: Cochrane review,"

British Journal of Cancer, vol. 96, no. 7, pp. 1025–1029, Apr. 2007, ISSN: 0007-0920. DOI: 10.1038/sj.bjc.6603600.

[262] W. Nijmeijer, "The evolution of clinical care and transmural medical pathways for frail older adults with hip fractures," English, Ph.D. dissertation, University of Twente, Netherlands, Jul. 2022, ISBN: 98-90-365-5376-6. DOI: 10.3990/1.9789036553766.

[263] E. W. de Bekker-Grob, B. Donkers, M. Bliemer, J. Coast, and J. Swait, "Towards Accurate Prediction of Healthcare Choices: The INTERSOCIAL Project," *The Patient - Patient-Centered Outcomes Research*, vol. 15, no. 5, pp. 509–512, Sep. 2022, ISSN: 1178-1661. DOI: 10.1007/s40271-022-00593-9.

[264] E. Lancsar and J. Swait, "Reconceptualising the External Validity of Discrete Choice Experiments," *PharmacoEconomics*, vol. 32, no. 10, pp. 951–965, Oct. 2014, ISSN: 1170-7690, 1179-2027. DOI: 10.1007/s40273-014-0181-7.

[265] A. Pilnick and R. Dingwall, "On the remarkable persistence of asymmetry in doctor/patient interaction: A critical review," *Social Science & Medicine*, vol. 72, no. 8, pp. 1374–1382, Apr. 2011, ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2011.02.033.

[266] O. S. Lian, S. Nettleton, H. Grange, and C. Dowrick, ""I'm not the doctor; I'm just the patient": Patient agency and shared decision-making in naturally occurring primary care consultations," *Patient Education and Counseling*, vol. 105, no. 7, pp. 1996–2004, Jul. 2022, ISSN: 0738-3991. DOI: 10.1016/j.pec.2021.10.031.

[267] T. M. P. Nijdam, D. W. P. M. Laane, J. F. Spierings, H. J. Schuijt, D. P. J. Smeeing, and D. van der Velde, "Proxy-reported experiences of palliative, non-operative management of geriatric patients after a hip fracture: a qualitative study," *BMJ Open*, vol. 12, no. 8, e063007, Aug. 2022, ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2022-063007.

[268] K. M. Culhane, M. O'connor, D. Lyons, and G. Lyons, "Accelerometers in rehabilitation medicine for older adults," *Age and ageing*, vol. 34, no. 6, pp. 556–560, 2005. DOI: 10.1093/ageing/afi192.

[269] J. B. Bussmann, J. H. Tulen, E. C. van Herel, and H. J. Stam, "Quantification of physical activities by means of ambulatory accelerometry: A validation study," *Psychophysiology*, vol. 35, no. 5, pp. 488–496, 1998, ISSN: 1469-8986. DOI: 10.1017/S0048577298971153.

[270] J. B. J. Bussmann, W. L. J. Martens, J. H. M. Tulen, F. C. Schasfoort, H. J. G. van den Berg-Emons, and H. J. Stam, "Measuring daily behavior using ambulatory accelerometry: The Activity Monitor," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 3, pp. 349–356, Aug. 2001, ISSN: 0743-3808, 1532-5970. DOI: 10.3758/BF03195388.

[271] S.-C. Kim, M. J. Kim, N. Kim, J. H. Hwang, and G. C. Han, "Ambulatory balance monitoring using a wireless attachable three-axis accelerometer," *Journal of Vestibular Research: Equilibrium & Orientation*, vol. 23, no. 4-5, pp. 217–225, 2013, ISSN: 1878-6464. DOI: 10.3233/VES-130489.

[272] K. M. Culhane, G. M. Lyons, D. Hilton, P. A. Grace, and D. Lyons, "Long-term mobility monitoring of older adults using accelerometers in a clinical environment," *Clinical Rehabilitation*, vol. 18, no. 3, pp. 335–343, May 2004, ISSN: 0269-2155. DOI: 10.1191/0269215504cr734oa.

[273] J. J. Kavanagh and H. B. Menz, "Accelerometry: a technique for quantifying movement patterns during walking," *Gait & posture*, vol. 28, no. 1, pp. 1–15, 2008. DOI: 10.1016/j.gaitpost.2007.10.010.

[274] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Systems with Applications*, vol. 137, pp. 167–190, Dec. 2019, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.04.057.

[275] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, p. 1 550 147 716 665 520, Aug. 2016, ISSN: 1550-1329. DOI: 10.1177/1550147716665520.

[276] W. Bijnens, J. Aarts, A. Stevens, D. Ummels, and K. Meijer, "Optimization and validation of an adjustable activity classification algorithm for assessment of physical behavior in elderly," *Sensors*, vol. 19, no. 24, p. 5344, 2019. DOI: 10.3390/s19245344.

[277] K. Aminian, P. Robert, E. E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon, "Physical activity monitoring based on accelerometry: validation and comparison with video observation," *Medical & Biological Engineering & Computing*, vol. 37, no. 3, pp. 304–308, May 1999, ISSN: 1741-0444. DOI: 10.1007/BF02513304.

[278] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010. DOI: 10.3390/s100807772.

[279] C. V. C. Bouten, K. T. M. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE transactions on biomedical engineering*, vol. 44, no. 3, pp. 136–147, 1997. DOI: 10.1109/10.554760.

[280] P. Veltink, H. B. Bussmann, W. de Vries, W. L. Martens, and R. Van Lummel, "Detection of static and dynamic activities using uniaxial accelerometers," *IEEE Transactions on Rehabilitation Engineering*, vol. 4, no. 4, pp. 375–385, Dec. 1996, Conference Name: IEEE Transactions on Rehabilitation Engineering, ISSN: 1558-0024. DOI: 10.1109/86.547939.

[281] S. M. Patterson, D. S. Krantz, L. C. Montgomery, P. A. Deuster, S. M. Hedges, and L. E. Nebel, "Automated physical activity monitoring: Validation and comparison with physiological and self-report measures," *Psychophysiology*, vol. 30, no. 3, pp. 296–305, 1993. DOI: 10.1111/j.1469-8986.1993.tb03356.x.

[282] M. J. Mathie, A. C. Coster, N. H. Lovell, and B. G. Celler, "Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement," *Physiological measurement*, vol. 25, no. 2, R1, 2004. DOI: 10.1088/0967-3334/25/2/r01.

[283] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014. DOI: 10.3390/s140406474.

[284] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013, ISSN: 1553-877X. DOI: 10.1109/SURV.2012.110112.00192.

[285] M. Straczkiewicz, P. James, and J.-P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–15, Oct. 2021, ISSN: 2398-6352. DOI: 10.1038/s41746-021-00514-4.

[286] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015. DOI: 10.3390/s151229858.

[287] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey," in *23th International Conference on Architecture of Computing Systems 2010*, Feb. 2010, pp. 1–10.

[288] A. Ignatov, "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018. DOI: 10.1016/j.asoc.2017.09.027.

[289] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, *Deep Activity Recognition Models with Triaxial Accelerometers*, 2016. arXiv: 1511.04664 [cs.LG]. (visited on 08/12/2021).

[290] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE international conference on systems, man, and cybernetics*, IEEE, 2015, pp. 1488–1492. DOI: 10.1109/SMC.2015.263.

[291] N. Y. Hammerla, S. Halloran, and T. Ploetz, *Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables*, 2016. arXiv: 1604.08880 [cs.LG]. (visited on 08/12/2021).

[292] M. Inoue, S. Inoue, and T. Nishida, *Deep Recurrent Neural Network for Mobile Human Activity Recognition with High Throughput*, 2016. arXiv: 1611.03607 [cs.CV]. (visited on 08/12/2021).

[293] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th international conference on mobile computing, applications and services*, IEEE, 2014, pp. 197–205. DOI: 10.4108/icst.mobicase.2014.257786.

[294] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016. DOI: 10.3390/s16010115.

[295] A. K. Bourke, E. A. F. Ihlen, and J. L. Helbostad, "Development of a gold-standard method for the identification of sedentary, light and moderate physical activities in older adults: Definitions for video annotation," *Journal of Science and Medicine in Sport*, vol. 22, no. 5, pp. 557–561, May 2019, ISSN: 14402440. DOI: 10.1016/j.jsams.2018.11.011.

[296] F. R. Allen, E. Ambikairajah, N. H. Lovell, and B. G. Celler, "Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models," *Physiological Measurement*, vol. 27, no. 10, pp. 935–951, Oct. 2006, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/27/10/001.

[297] S. Chernbumroong, S. Cang, A. Atkins, and H. Yu, "Elderly activities recognition and classification for applications in assisted living," *Expert Systems with*

Applications, vol. 40, no. 5, pp. 1662–1674, Apr. 2013, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2012.09.004.

[298] Ç. B. Erdaş, I. Atasoy, K. Açıcı, and H. Oğul, "Integrating features for accelerometer-based activity recognition," *Procedia Computer Science*, vol. 98, pp. 522–527, 2016, ISSN: 18770509. DOI: 10.1016/j.procs.2016.09.070.

[299] P. Gupta and T. Dallas, "Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1780–1786, Jun. 2014, ISSN: 1558-2531. DOI: 10.1109/TBME.2014.2307069.

[300] M. Janidarmian, A. Roshan Fekr, K. Radecka, and Z. Zilic, "A comprehensive analysis on wearable acceleration sensors in human activity recognition," *Sensors*, vol. 17, no. 3, p. 529, 2017. DOI: 10.3390/s17030529.

[301] S. L. Lau and K. David, "Movement recognition using the accelerometer in smartphones," in *2010 Future Network & Mobile Summit*, IEEE, 2010, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/document/5722356 (visited on 02/02/2022).

[302] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity Recognition Using a Single Accelerometer Placed at the Wrist or Ankle," *Medicine & Science in Sports & Exercise*, vol. 45, no. 11, pp. 2193–2203, Nov. 2013, ISSN: 0195-9131. DOI: 10.1249/MSS.0b013e31829736d6.

[303] N. Pannurat, S. Thiemjarus, E. Nantajeewarawat, and I. Anantavrasilp, "Analysis of optimal sensor positions for activity classification and application on a different data collection scenario," *Sensors*, vol. 17, no. 4, p. 774, 2017. DOI: 10.3390/s17040774.

[304] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, Pittsburgh, PA, vol. 5, 2005, pp. 1541–1546. [Online]. Available: https://www.aaai.org/Papers/IAAI/2005/IAAI05-013.pdf (visited on 08/11/2021).

[305] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, Springer, 2004, pp. 1–17. DOI: 10.1007/978-3-540-24646-6_1.

[306] A. Bayat, M. Pomplun, and D. A. Tran, "A Study on Human Activity Recognition Using Accelerometer Data from Smartphones," *Procedia Computer Science*, vol. 34, pp. 450–457, 2014, ISSN: 18770509. DOI: 10.1016/j.procs.2014.07.009.

[307] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, S. McClean, and D. Finlay, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors*, vol. 13, no. 7, pp. 9183–9200, 2013. DOI: 10.3390/s130709183.

[308] D. O. Olguın and A. S. Pentland, "Human activity recognition: Accuracy across common locations for wearable sensors," in *Proceedings of 2006 10th IEEE international symposium on wearable computers, Montreux, Switzerland*, Citeseer, 2006, pp. 11–14. [Online]. Available: https://hd.media.mit.edu/tech-reports/TR-603.pdf (visited on 09/02/2021).

[309] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors—a review of

classification techniques," *Physiological measurement*, vol. 30, no. 4, R1, 2009. DOI: 10.1088/0967-3334/30/4/R01.

[310] C. Xia and Y. Sugiura, "Wearable Accelerometer Optimal Positions for Human Motion Recognition," in *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, 2020, pp. 19–20. DOI: 10.1109/LifeTech48969.2020.1570618961.

[311] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE transactions on biomedical circuits and systems*, vol. 5, no. 4, pp. 320–329, 2011. DOI: 10.1109/TBCAS.2011.2160540.

[312] M. Arif and A. Kattan, "Physical activities monitoring using wearable acceleration sensors attached to the body," *PloS one*, vol. 10, no. 7, e0130851, 2015. DOI: 10.1371/journal.pone.0130851.

[313] S. Pirttikangas, K. Fujinami, and T. Nakajima, "Feature selection and activity recognition from wearable sensors," in *International symposium on ubiquitious computing systems*, Springer, 2006, pp. 516–527. DOI: 10.1007/11890348_39.

[314] A. M. Khan, Y.-K. Lee, S. Lee, and T.-S. Kim, "Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly," *Medical & biological engineering & computing*, vol. 48, no. 12, pp. 1271–1279, 2010. DOI: 10.1007/s11517-010-0701-3.

[315] K. Kunze and P. Lukowicz, "Sensor placement variations in wearable activity recognition," *IEEE Pervasive Computing*, vol. 13, no. 4, pp. 32–41, 2014. DOI: 10.1109/MPRV.2014.73.

[316] I. Orha and S. Oniga, "Study regarding the optimal sensors placement on the body for human activity recognition," in *2014 IEEE 20th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2014, pp. 203–206. DOI: 10.1109/SIITME.2014.6967028.

[317] C. Bai, A. A. Wanigatunga, S. Saldana, R. Casanova, T. M. Manini, and M. T. Mardini, "Are Machine Learning Models on Wrist Accelerometry Robust against Differences in Physical Performance among Older Adults?" *Sensors (Basel, Switzerland)*, vol. 22, no. 8, p. 3061, Apr. 2022, ISSN: 1424-8220. DOI: 10.3390/s22083061.

[318] M. Janidarmian, A. Roshan Fekr, K. Radecka, Z. Zilic, and L. Ross, "Analysis of Motion Patterns for Recognition of Human Activities," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, ser. MOBIHEALTH'15, Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Dec. 2015, pp. 68–72, ISBN: 978-1-63190-088-4. DOI: 10.4108/eai.14-10-2015.2261719.

[319] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A Quantitative Comparison of Overlapping and Non-Overlapping Sliding Windows for Human Activity Recognition Using Inertial Sensors," *Sensors*, vol. 19, no. 22, p. 5026, Jan. 2019, ISSN: 1424-8220. DOI: 10.3390/s19225026.

[320] M. Sun and J. O. Hill, "A method for measuring mechanical work and work efficiency during human activities," *Journal of biomechanics*, vol. 26, no. 3, pp. 229–241, 1993. DOI: 10.1016/0021-9290(93)90361-h.

[321] N. Twomey, T. Diethe, X. Fafoutis, A. Elsts, R. McConville, P. Flach, and I. Craddock, "A comprehensive study of activity recognition using accelerometers," in *Informatics*, Multidisciplinary Digital Publishing Institute, vol. 5, 2018, p. 27. DOI: 10.3390/informatics5020027.

[322] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2008. DOI: 10.1109/TBME.2008.2006190.

[323] N. Wang, E. Ambikairajah, N. H. Lovell, and B. G. Celler, "Accelerometry Based Classification of Walking Patterns Using Time-frequency Analysis," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2007, pp. 4899–4902. DOI: 10.1109/IEMBS.2007.4353438.

[324] T. Huynh and B. Schiele, "Analyzing features for activity recognition," in *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, 2005, pp. 159–163. DOI: 10.1145/1107548.1107591.

[325] G. Wang, Q. Li, L. Wang, W. Wang, M. Wu, and T. Liu, "Impact of Sliding Window Length in Indoor Human Motion Modes and Pose Pattern Recognition Based on Smartphone Sensors," *Sensors*, vol. 18, no. 6, p. 1965, Jun. 2018, ISSN: 1424-8220. DOI: 10.3390/s18061965.

[326] A. Dehghani, T. Glatard, and E. Shihab, *Subject Cross Validation in Human Activity Recognition*, Apr. 2019. DOI: 10.48550/ARXIV.1904.02666.

[327] T. Plötz, N. Y. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*, ser. IJCAI'11, Barcelona, Catalonia, Spain: AAAI Press, Jul. 2011, pp. 1729–1734, ISBN: 978-1-57735-514-4. DOI: 10.5555/2283516.2283683.

[328] J. Baek, G. Lee, W. Park, and B.-J. Yun, "Accelerometer signal processing for user activity detection," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, vol. 3215, 2004, pp. 610–617, ISBN: 978-3-540-23205-6. DOI: 10.1007/978-3-540-30134-9_82.

[329] S. G. Trost, Y. Zheng, and W.-K. Wong, "Machine learning for activity recognition: hip versus wrist data," *Physiological Measurement*, vol. 35, no. 11, pp. 2183–2189, Nov. 2014, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/35/11/2183.

[330] M. Bennasar, B. A. Price, D. Gooch, A. K. Bandara, and B. Nuseibeh, "Significant Features for Human Activity Recognition Using Tri-Axial Accelerometers," *Sensors*, vol. 22, no. 19, p. 7482, Jan. 2022, Number: 19 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s22197482.

[331] J. Chen, Y. Sun, and S. Sun, "Improving Human Activity Recognition Performance by Data Fusion and Feature Engineering," *Sensors (Basel, Switzerland)*, vol. 21, no. 3, p. 692, Jan. 2021, ISSN: 1424-8220. DOI: 10.3390/s21030692.

[332] M. J. Mathie, A. C. F. Coster, N. H. Lovell, and B. G. Celler, "Detection of daily physical activities using a triaxial accelerometer," *Medical and Biological Engineering and Computing*, vol. 41, no. 3, pp. 296–301, 2003. DOI: 10.1007/BF02348434.

[333] J. Seo, Y. Chiang, T. H. Laine, and A. M. Khan, "Step counting on smartphones using advanced zero-crossing and linear regression," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, 2015, pp. 1–7. DOI: 10.1145/2701126.2701223.

[334] B. Nham, K. Siangliulue, and S. Yeung, "Predicting mode of transport from iphone accelerometer data," *Machine Learning Final Projects, Stanford University*, 2008. [Online]. Available: https://www.semanticscholar.org/paper/Predicting-Mode-of-Transport-from-iPhone-Data-Nham-Siangliulue/111a8a8faf04e754dcef1f0d4701a2d6af36ce8b (visited on 08/11/2021).

[335] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2213–2220, Dec. 2008, ISSN: 01678655. DOI: 10.1016/j.patrec.2008.08.002.

[336] T. Tamura, M. Sekine, M. Ogawa, T. Togawa, and Y. Fukui, "Classification of Acceleration Waveforms during Walking by Wavelet Transform," *Methods of Information in Medicine*, vol. 36, no. 04/05, pp. 356–359, Oct. 1997, ISSN: 0026-1270, 2511-705X. DOI: 10.1055/s-0038-1636855.

[337] M. Sekine, T. Tamura, T. Fujimoto, and Y. Fukui, "Classification of walking pattern using acceleration waveform in elderly people," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, vol. 2, Jul. 2000, pp. 1356–1359. DOI: 10.1109/IEMBS.2000.897990.

[338] W. Cochran, J. Cooley, D. Favin, H. Helms, R. Kaenel, W. Lang, G. Maling, D. Nelson, C. Rader, and P. Welch, "What is the fast Fourier transform?" *Proceedings of the IEEE*, vol. 55, no. 10, pp. 1664–1674, Oct. 1967, ISSN: 1558-2256. DOI: 10.1109/PROC.1967.5957.

[339] C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, Jan. 1998, ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

[340] C. L. Webber and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *Journal of Applied Physiology*, vol. 76, no. 2, pp. 965–973, Feb. 1994, ISSN: 8750-7587. DOI: 10.1152/jappl.1994.76.2.965.

[341] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. P. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645–662, 2010. DOI: 10.1007/s00779-010-0293-9.

[342] Q. T. Huynh, U. D. Nguyen, K. T. Liem, and B. Q. Tran, "Detection of activities daily living and falls using combination accelerometer and gyroscope," in *5th International Conference on Biomedical Engineering in Vietnam*, Springer, 2015, pp. 184–189. DOI: 10.1007/978-3-319-11776-8_45.

[343] Y.-P. Chen, J.-Y. Yang, S.-N. Liou, G.-Y. Lee, and J.-S. Wang, "Online classifier construction algorithm for human activity detection using a tri-axial accelerometer,"

Applied Mathematics and Computation, vol. 205, no. 2, pp. 849–860, 2008. DOI: 10.1016/j.amc.2008.05.099.

[344] N. A. Capela, E. D. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PloS one*, vol. 10, no. 4, e0124414, 2015. DOI: 10.1371/journal.pone.0124414.

[345] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, Sep. 2018, ISSN: 15320464. DOI: 10.1016/j.jbi.2018.07.014.

[346] A. Davoudi, M. T. Mardini, D. Nelson, F. Albinali, S. Ranka, P. Rashidi, and T. M. Manini, "The Effect of Sensor Placement and Number on Physical Activity Recognition and Energy Expenditure Estimation in Older Adults: Validation Study," *JMIR mHealth and uHealth*, vol. 9, no. 5, e23681, May 2021, ISSN: 2291-5222. DOI: 10.2196/23681.

[347] Y. Zhang, S. Markovic, I. Sapir, R. C. Wagenaar, and T. D. C. Little, "Continuous functional activity monitoring based on wearable tri-axial accelerometer and gyroscope," in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, IEEE, 2011, pp. 370–373. [Online]. Available: https://ieeexplore.ieee.org/document/6038832 (visited on 02/02/2022).

[348] L. Verlaan, S. Bolink, S. Van Laarhoven, M. Lipperts, I. Heyligers, B. Grimm, and R. Senden, "Accelerometer-based Physical Activity Monitoring in Patients with Knee Osteoarthritis: Objective and Ambulatory Assessment of Actual Physical Activity During Daily Life Circumstances," *The Open Biomedical Engineering Journal*, vol. 9, pp. 157–163, Jul. 2015, ISSN: 1874-1207. DOI: 10.2174/1874120701509010157.

[349] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1254, 2018, ISSN: 1942-4795. DOI: 10.1002/widm.1254.

[350] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003, ISSN: 1532-4435. DOI: 10.5555/944919.944968.

[351] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter Methods for Feature Selection – A Comparative Study," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 178–187, ISBN: 978-3-540-77226-2. DOI: 10.1007/978-3-540-77226-2_19.

[352] V. L. Tang, R. Sudore, I. S. Cenzer, W. J. Boscardin, A. Smith, C. Ritchie, M. Wallhagen, E. Finlayson, L. Petrillo, and K. Covinsky, "Rates of Recovery to Pre-Fracture Function in Older Persons with Hip Fracture: an Observational Study," *Journal of General Internal Medicine*, vol. 32, no. 2, pp. 153–158, Feb. 2017, ISSN: 0884-8734, 1525-1497. DOI: 10.1007/s11606-016-3848-2. [Online]. Available: http://link.springer.com/10.1007/s11606-016-3848-2.

[353] J. B. Talkowski, E. J. Lenze, M. C. Munin, C. Harrison, and J. S. Brach, "Patient Participation and Physical Activity During Rehabilitation and Future Functional Outcomes in Patients After Hip Fracture," *Archives of Physical Medicine and Rehabilitation*, vol. 90, no. 4, pp. 618–622, Apr. 2009, ISSN: 0003-9993. DOI: 10.1016/j.apmr.2008.10.024.

[354] A. J. Vochteloo, S. Moerman, W. E. Tuinebreijer, A. B. Maier, M. R. de Vries, R. M. Bloem, R. G. Nelissen, and P. Pilot, "More than half of hip fracture patients do not regain mobility in the first postoperative year: Mobility and hip fracture," *Geriatrics & Gerontology International*, vol. 13, no. 2, pp. 334–341, Apr. 2013, ISSN: 14441586. DOI: 10.1111/j.1447-0594.2012.00904.x.

[355] J. A. Pasco, K. M. Sanders, F. M. Hoekstra, M. J. Henry, G. C. Nicholson, and M. A. Kotowicz, "The human cost of fracture," *Osteoporosis International*, vol. 16, no. 12, pp. 2046–2052, Dec. 2005, ISSN: 0937-941X, 1433-2965. DOI: 10.1007/s00198-005-1997-y.

[356] T. Hida, A. Harada, S. Imagama, and N. Ishiguro, "Managing Sarcopenia and Its Related-Fractures to Improve Quality of Life in Geriatric Populations," *Aging and Disease*, vol. 5, no. 4, pp. 226–237, Nov. 2013, ISSN: 2152-5250. DOI: 10.14336/AD.2014.0500226.

[357] D. Taylor, "Physical activity is medicine for older adults," *Postgraduate Medical Journal*, vol. 90, no. 1059, pp. 26–32, Jan. 2014, ISSN: 0032-5473, 1469-0756. DOI: 10.1136/postgradmedj-2012-131366.

[358] W. J. Chodzko-Zajko, D. N. Proctor, M. A. Fiatarone Singh, C. T. Minson, C. R. Nigg, G. J. Salem, and J. S. Skinner, "Exercise and Physical Activity for Older Adults," *Medicine & Science in Sports & Exercise*, vol. 41, no. 7, pp. 1510–1530, Jul. 2009, ISSN: 0195-9131. DOI: 10.1249/MSS.0b013e3181a0c95c.

[359] C. Daskalopoulou, B. Stubbs, C. Kralj, A. Koukounari, M. Prince, and A. M. Prina, "Physical activity and healthy ageing: A systematic review and meta-analysis of longitudinal cohort studies," *Ageing Research Reviews*, vol. 38, pp. 6–17, Sep. 2017, ISSN: 1568-1637. DOI: 10.1016/j.arr.2017.06.003.

[360] W. Hollmann, H. K. Strüder, C. V. Tagarakis, and G. King, "Physical activity and the elderly," *European Journal of Cardiovascular Prevention & Rehabilitation*, vol. 14, no. 6, pp. 730–739, Dec. 2007, Publisher: SAGE Publications, ISSN: 1741-8267. DOI: 10.1097/HJR.0b013e32828622f9.

[361] J. F. Fries, "Physical activity, the compression of morbidity, and the health of the elderly.," *Journal of the Royal Society of Medicine*, vol. 89, no. 2, pp. 64–68, Feb. 1996, ISSN: 0141-0768.

[362] "Sarcopenia: An Undiagnosed Condition in Older Adults. Current Consensus Definition: Prevalence, Etiology, and Consequences. International Working Group on Sarcopenia," vol. 12, pp. 249–256, May 2011, ISSN: 1525-8610. DOI: 10.1016/j.jamda.2011.01.003.

[363] B. Resnick, E. Galik, M. Boltz, W. Hawkes, M. Shardell, D. Orwig, and J. Magaziner, "Physical Activity in the Post-Hip-Fracture Period," *Journal of aging and physical activity*, vol. 19, no. 4, p. 373, Oct. 2011, Publisher: NIH Public Access. DOI: 10.1123/japa.19.4.373.

[364] L. Fleig, M. M. McAllister, P. Brasher, W. L. Cook, P. Guy, J. H. Puyat, K. M. Khan, H. A. McKay, and M. C. Ashe, "Sedentary Behavior and Physical Activity

Patterns in Older Adults After Hip Fracture: A Call to Action," *Journal of Aging and Physical Activity*, vol. 24, no. 1, pp. 79–84, Jan. 2016, ISSN: 1543-267X. DOI: 10.1123/japa.2015-0013.

[365] C. E. Matthews, B. E. Ainsworth, R. W. Thompson, and D. R. Bassett, "Sources of variance in daily physical activity levels as measured by an accelerometer," *Medicine and Science in Sports and Exercise*, vol. 34, no. 8, pp. 1376–1381, Aug. 2002, ISSN: 0195-9131. DOI: 10.1097/00005768-200208000-00021.

[366] J. Annegarn, M. A. Spruit, N. H. M. K. Uszko-Lencer, S. Vanbelle, H. H. C. M. Savelberg, A. M. W. J. Schols, E. F. M. Wouters, and K. Meijer, "Objective Physical Activity Assessment in Patients With Chronic Organ Failure: A Validation Study of a New Single-Unit Activity Monitor," *Archives of Physical Medicine and Rehabilitation*, vol. 92, no. 11, Nov. 2011, ISSN: 0003-9993, 1532-821X. DOI: 10.1016/j.apmr.2011.06.021.

[367] M. Iosa, A. Fusco, G. Morone, and S. Paolucci, "Development and decline of upright gait stability," *Frontiers in aging neuroscience*, vol. 6, p. 14, 2014. DOI: 10.3389/fnagi.2014.00014.

[368] M. B. Del Rosario, K. Wang, J. Wang, Y. Liu, M. Brodie, K. Delbaere, N. H. Lovell, S. R. Lord, and S. J. Redmond, "A comparison of activity classification in younger and older cohorts using a smartphone," *Physiological Measurement*, vol. 35, no. 11, pp. 2269–2286, Nov. 2014, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/35/11/2269.

[369] K.-J. Lee, S.-H. Um, and Y.-H. Kim, "Postoperative Rehabilitation after Hip Fracture: A Literature Review," *Hip & Pelvis*, vol. 32, no. 3, pp. 125–131, Sep. 2020, ISSN: 2287-3260. DOI: 10.5371/hp.2020.32.3.125.

[370] A. Logacjov, K. Bach, A. Kongsvold, H. B. Bårdstu, and P. J. Mork, "HARTH: A Human Activity Recognition Dataset for Machine Learning," *Sensors*, vol. 21, no. 23, p. 7853, Jan. 2021, ISSN: 1424-8220. DOI: 10.3390/s21237853.

[371] R. Younes, M. Jones, and T. L. Martin, "Classifier for Activities with Variations," *Sensors*, vol. 18, no. 10, p. 3529, Oct. 2018, ISSN: 1424-8220. DOI: 10.3390/s18103529.

[372] K. Akila and S. Chitrakala, "Highly refined human action recognition model to handle intraclass variability & interclass similarity," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20 877–20 894, Aug. 2019, ISSN: 1380-7501, 1573-7721. DOI: 10.1007/s11042-019-7392-z.

[373] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, no. 3, Jan. 2014, ISSN: 0360-0300. DOI: 10.1145/2499621.

[374] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A Review of Human Activity Recognition Methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015, ISSN: 2296-9144. DOI: 10.3389/frobt.2015.00028.

[375] B. Barshan and A. Yurtman, "Investigating Inter-Subject and Inter-Activity Variations in Activity Recognition Using Wearable Motion Sensors," *The Computer Journal*, vol. 59, no. 9, pp. 1345–1362, Sep. 2016, ISSN: 1460-2067. DOI: 10.1093/comjnl/bxv093.

[376] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "On the Personalization of Classification Models for Human Activity Recognition," *IEEE Access*, vol. 8,

pp. 32 066–32 079, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2973425.

[377] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Fusion of Smartphone Motion Sensors for Physical Activity Recognition," *Sensors*, vol. 14, no. 6, pp. 10 146–10 176, Jun. 2014, ISSN: 1424-8220. DOI: 10.3390/s140610146.

[378] ——, "Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors," *Sensors*, vol. 16, no. 4, p. 426, Apr. 2016, ISSN: 1424-8220. DOI: 10.3390/s16040426.

[379] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac60214a047.

[380] R. W. Schafer, "On the frequency-domain properties of Savitzky-Golay filters," in *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, Jan. 2011, pp. 54–59. DOI: 10.1109/DSP-SPE.2011.5739186.

[381] H. Lee, H. Lee, and M. Whang, "An Enhanced Method to Estimate Heart Rate from Seismocardiography via Ensemble Averaging of Body Movements at Six Degrees of Freedom," *Sensors*, vol. 18, no. 1, p. 238, Jan. 2018, ISSN: 1424-8220. DOI: 10.3390/s18010238.

[382] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, Mar. 2011, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2010.09.013.

[383] F. Petitjean and P. Gançarski, "Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment," *Theoretical Computer Science*, vol. 414, no. 1, pp. 76–91, Jan. 2012, ISSN: 03043975. DOI: 10.1016/j.tcs.2011.09.029.

[384] R. S. L. Chan, P. Gordon, and M. R. Smith, "Evaluation of Dynamic Time Warp Barycenter Averaging (DBA) for its Potential in Generating a Consensus Nanopore Signal for Genetic and Epigenetic Sequences," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, ISSN: 1558-4615, Jul. 2018, pp. 2821–2824. DOI: 10.1109/EMBC.2018.8512873.

[385] Z. Zhao, H. Fang, S. Williams, S. D. Relton, J. Alty, A. J. Casson, and D. C. Wong, "Time series clustering to examine presence of decrement in Parkinson's finger-tapping bradykinesia," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, ISSN: 2694-0604, Jul. 2020, pp. 780–783. DOI: 10.1109/EMBC44109.2020.9175638.

[386] T. Hachaj, K. Koptyra, and M. R. Ogiela, "Averaging of motion capture recordings for movements' templates generation," *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 30 353–30 380, Dec. 2018, ISSN: 1380-7501, 1573-7721. DOI: 10.1007/s11042-018-6137-8.

[387] G. B. Papini, P. Fonseca, X. L. Aubert, S. Overeem, J. W. M. Bergmans, and R. Vullings, "Photoplethysmography beat detection and pulse morphology quality assessment for signal reliability estimation," *Annual International Conference*

of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2017, pp. 117–120, Jul. 2017, ISSN: 2694-0604. DOI: 10.1109/EMBC.2017.8036776.

[388] S. Seto, W. Zhang, and Y. Zhou, *Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition*, arXiv:1512.06747 [cs], Dec. 2015. DOI: 10.48550/arXiv.1512.06747.

[389] L. Etienne, T. Devogele, M. Buchin, and G. McArdle, "Trajectory Box Plot: a new pattern to summarize movements," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 835–853, May 2016, ISSN: 1365-8816. DOI: 10.1080/13658816.2015.1081205.

[390] "Dynamic Time Warping," in *Information Retrieval for Music and Motion*, M. Müller, Ed., Berlin, Heidelberg: Springer, 2007, pp. 69–84, ISBN: 978-3-540-74048-3. DOI: 10.1007/978-3-540-74048-3_4.

[391] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[392] M. T. Uddin, M. M. Billah, and M. F. Hossain, "Random forests based recognition of human activities and postural transitions on smartphone," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, IEEE, 2016, pp. 250–255.

[393] L. Chen, S. Fan, V. Kumar, and Y. Jia, "A method of human activity recognition in transitional period," *Information*, vol. 11, no. 9, p. 416, 2020.

[394] A. Elsts, R. McConville, X. Fafoutis, N. Twomey, R. J. Piechocki, R. Santos-Rodriguez, and I. Craddock, "On-Board Feature Extraction from Acceleration Data for Activity Recognition.," in *EWSN*, 2018, pp. 163–168. [Online]. Available: https://dl.acm.org/doi/10.5555/3234847.3234868 (visited on 08/11/2021).

[395] C. Wallisch, D. Dunkler, G. Rauch, R. de Bin, and G. Heinze, "Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling," *Statistics in Medicine*, vol. 40, no. 2, pp. 369–381, 2021, ISSN: 1097-0258. DOI: 10.1002/sim.8779.

[396] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Information Fusion*, vol. 52, pp. 1–12, Dec. 2019, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2018.11.008.

[397] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, "Testing Different Ensemble Configurations for Feature Selection," *Neural Processing Letters*, vol. 46, no. 3, pp. 857–880, Dec. 2017, ISSN: 1370-4621, 1573-773X. DOI: 10.1007/s11063-017-9619-1.

[398] P. Drotár, M. Gazda, and L. Vokorokos, "Ensemble feature selection using election methods and ranker clustering," *Information Sciences*, vol. 480, pp. 365–380, Apr. 2019, ISSN: 0020-0255. DOI: 10.1016/j.ins.2018.12.033.

[399] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognition*, vol. 45, no. 1, pp. 531–539, Jan. 2012, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.06.006.

[400] A. Ben Brahim and M. Limam, "Ensemble feature selection for high dimensional data: a new method and a comparative study," *Advances in Data Analysis and Classification*, vol. 12, no. 4, pp. 937–952, Dec. 2018, ISSN: 1862-5347, 1862-5355. DOI: 10.1007/s11634-017-0285-y.

[401] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, May 2007, ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-006-0040-8.

[402] F. Bunea, Y. She, H. Ombao, A. Gongvatana, K. Devlin, and R. Cohen, "Penalized least squares regression methods and applications to neuroimaging," *NeuroImage*, vol. 55, no. 4, pp. 1519–1527, Apr. 2011, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2010.12.028.

[403] E. H. Blackstone, "Breaking down barriers: Helpful breakthrough statistical methods you need to understand better," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 122, no. 3, pp. 430–439, Sep. 2001, ISSN: 00225223. DOI: 10.1067/mtc.2001.117536.

[404] A. M. Salarbaks, R. Lindeboom, and W. Nijmeijer, "Pneumonia in hospitalized elderly hip fracture patients: the effects on length of hospital-stay, in-hospital and thirty-day mortality and a search for potential predictors," *Injury*, vol. 51, no. 8, Aug. 2020, ISSN: 0020-1383, 1879-0267. DOI: 10.1016/j.injury.2020.05.017.

[405] C. Ding and H. Peng, "Minimum Redundancy Feature Selection From Microarray Gene Expression Data," *Journal of Bioinformatics and Computational Biology*, Nov. 2011, Publisher: Imperial College Press. DOI: 10.1142/S0219720005001004.

[406] W.-Y. Loh, "Regression Trees With Unbiased Variable Selection and Interaction Detection," *Statistica Sinica*, vol. 12, pp. 361–386, Apr. 2002.

[407] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[408] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj035.

[409] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, ISSN: 00457906. DOI: 10.1016/j.compeleceng.2013.11.024.

[410] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, Feb. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr709.

[411] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science*, vol. 302, no. 5643, pp. 249–255, Oct. 2003. DOI: 10.1126/science.1087447.

[412] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, May 2006, Number: 5 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt1203.

[413] B. Venkatesh and J. Anuradha, "A Review of Feature Selection and Its Methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, Mar. 2019. DOI: 10.2478/cait-2019-0001.

[414] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning," in. Springer, 2013, vol. 112, ch. 10, pp. 373–413. [Online]. Available: https://link.springer.com/content/pdf/10.1007%5C%2F978-1-4614-7138-7.pdf (visited on 06/06/2021).

[415] O. Banos, M. A. Toth, M. Damas, H. Pomares, and I. Rojas, "Dealing with the Effects of Sensor Displacement in Wearable Activity Recognition," *Sensors (Basel, Switzerland)*, vol. 14, no. 6, pp. 9995–10 023, Jun. 2014, ISSN: 1424-8220. DOI: 10.3390/s140609995.

[416] E. B. Gausden, D. Sin, A. E. Levack, L. E. Wessel, G. Moloney, J. M. Lane, and D. G. Lorich, "Gait Analysis After Intertrochanteric Hip Fracture: Does Shortening Result in Gait Impairment?" *Journal of Orthopaedic Trauma*, vol. 32, no. 11, pp. 554–558, Nov. 2018, ISSN: 1531-2291. DOI: 10.1097/BOT.0000000000001283.

[417] J. Houck, J. Kneiss, S. V. Bukata, and J. E. Puzas, "Analysis of vertical ground reaction force variables during a Sit to Stand task in participants recovering from a hip fracture," *Clinical Biomechanics*, vol. 26, no. 5, pp. 470–476, Jun. 2011, ISSN: 0268-0033. DOI: 10.1016/j.clinbiomech.2010.12.004.

[418] A. Arcelus, C. L. Herry, R. A. Goubran, F. Knoefel, H. Sveistrup, and M. Bilodeau, "Determination of Sit-to-Stand Transfer Duration Using Bed and Floor Pressure Sequences," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 10, pp. 2485–2492, Oct. 2009, Conference Name: IEEE Transactions on Biomedical Engineering, ISSN: 1558-2531. DOI: 10.1109/TBME.2009.2026733.

[419] M. Goffredo, M. Schmid, S. Conforto, M. Carli, A. Neri, and T. D'Alessio, "Markerless Human Motion Analysis in Gauss–Laguerre Transform Domain: An Application to Sit-To-Stand in Young and Elderly People," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 207–216, Mar. 2009, Conference Name: IEEE Transactions on Information Technology in Biomedicine, ISSN: 1558-0032. DOI: 10.1109/TITB.2008.2007960.

[420] A. Salarian, H. Russmann, F. J. G. Vingerhoets, P. R. Burkhard, and K. Aminian, "Ambulatory Monitoring of Physical Activities in Patients With Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2296–2299, Dec. 2007, Conference Name: IEEE Transactions on Biomedical Engineering, ISSN: 1558-2531. DOI: 10.1109/TBME.2007.896591.

[421] B. Najafi, K. Aminian, F. Loew, Y. Blanc, and P. Robert, "Measurement of stand-sit and sit-stand transitions using a miniature gyroscope and its application in fall risk evaluation in the elderly," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 8, pp. 843–851, Aug. 2002, Conference Name: IEEE Transactions on Biomedical Engineering, ISSN: 1558-2531. DOI: 10.1109/TBME.2002.800763.

[422] D. Pfeufer, C. Grabmann, S. Mehaffey, A. Keppler, W. Böcker, C. Kammerlander, and C. Neuerburg, "Weight bearing in patients with femoral neck fractures compared to pertrochanteric fractures: A postoperative gait analysis," *Injury*, vol. 50, no. 7, pp. 1324–1328, Jul. 2019, ISSN: 0020-1383. DOI: 10.1016/j.injury.2019.05.008.

APPENDIX A

Meta-Analysis Supplements

A.1 Study Selection and Characteristics

Table A.1: Overview of the database search query used for the systematic search of articles.

Term [Older adult]		Term [Hip fracture]		Term [Mortality]		Term [Predictor]
elderly OR *octogenarian** OR *nonagenarian** OR *"old* adult*"*	AND	*"hip fracture*"* OR *"proximal femure fracture*"* OR *"neck of femur fracture*"* OR *"cervical fracture*"* OR *"femoral neck fracture*"* OR *"trochanteric fracture*"* OR *"intertrochanteric fracture*"* OR *"subtrochanteric fracture*"* OR *"pertrochanteric fracture*"*	AND	*"early mortality"* OR *"inpatient mortality"* OR *"in-hospital mortality"* OR *"30-day mortality"* OR *"1-year mortality"* OR *"postoperative mortality"* OR *"post-operative mortality"* OR *"postoperative mortality"* OR *"postsurgical mortality"* OR *"post-surgical mortality"* OR *death* OR *surviv**	AND	*"risk factor*"* OR *"preoperative risk factor*"* OR *"pre-operative risk factor*"* OR *"presurgical risk factor*"* OR *"pre-surgical risk factor*"* OR *predict** OR *"preoperative predict*"* OR *"pre-operative predict*"* OR *"presurgical predict*"* OR *"pre-surgical predict*"*

Table A.2: Exclusion criteria used for the literature screening procedure.

Exclusion criteria	Explanation
E1 - The article does not report on primary data.	The article is a literature review or a systematic review.
E2 - The article describes an unrepresentative population.	The described study population has a mean or median age below 70 years, solely comprises a single gender, or solely describes a periprosthetic hip fracture population.
E3 - The article does not report on preoperative predictors.	The reported risk factors for mortality do not provide predictive information prior to surgery.
E4 - The article does not report independent risk factors.	The study does not control for confounding using multiple (Cox or logistic) regression analysis.
E5 - The article does not report the statistics of interest.	The odds and hazard ratios are not reported, or they are reported without 95% confidence intervals.
E6 - The article does not (clearly) report predictors of mortality.	Mortality is not an outcome, or it is reported as a composite score of multiple adverse events.
E7 - The article does not report on mortality within one year.	The lower a priori life expectancy of older adults decreases the predictive value of long-term studies.
E8 - No full access to the article can be acquired.	

Table A.3: Characteristics of studies included in the systematic review. The fracture types were categorised as femoral neck (FN), displaced femoral neck (DFN), undisplaced femoral neck (UFN), trochanteric (T), and other (O) fractures. The surgical procedures were categorised as internal fixation (IF), arthroplasty (A), hemiarthroplasty (HA), total hip arthroplasty (THA), and other (O). The risk measures were described with either an odds ratio (OR) or hazard ratio (HR).

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
Aharonoff (1997) [107]	United States	Retrospective	612	80.1%	≥65	FN: 53%, T: 47%	IF, A	1-year (12.7%)	HR
Aldebeyan (2017) [172]	United States, Canada	Retrospective	10,117	69.7%	Mean: 80.5	-	IF: 78.8%, HA: 18.6%, THA: 2.6%	30-day (-)	OR
Adunsky (2012) [106]	Israel	Retrospective	1,114	21.9%	Mean: 82.4	FN: 38.7%, T: 61.3%	IF: 60.7%, A: 39.3%	3-month (4.2%), 1-year (10.6%)	HR
Ariza-Vega (2015) [108]	Spain	Prospective	275	78.5%	Mean: 81.4	FN: 46.9%, T: 52.0%, O: 1.1%	IF: 65.1%, HA: 33.8%, O: 1.1%	1-year (21%)	HR
Baidoo (2021) [145]	Ghana	Prospective	76	52.6%	Mean: 75.8	FN: 50.0%, T: 50.0%	-	6-month (13.2%), 1-year (19.7%), 4-year (27.6%)	HR
Bell (2016) [153]	Australia	Prospective	322	71.4%	Median: 83.4	DFN: 34%, UFN: 19%, T: 46.1%, O: 0.9%	IF: 62.4%, HA: 30.7%, THA: 6.9%	1-year (23.9%)	OR
Bellelli (2012) [109]	Italy	Retrospective	390	81.8%	Mean: 83.7	FN: 46.2%, T: 53.8%	IF: 59.0%, A: 41.0%	1-year (-)	HR
Belmont (2014) [14]	United States	Retrospective	44,419	61.6%	Mean: 72.7	FN: 63.5%, T: 36.5%	IF, A	Inpatient (4.5%)	OR
Björkelund (2009) [110]	Sweden	Retrospective	428	72.9%	Mean: 82.5	DFN: 34.1%, UFN: 18.0%, T: 31.8%, O: 16.1%	IF: 70%, HA: 27%, THA: 3%	4-month (13.5%)	OR
Bliemel (2016) [154]	Germany	Prospective	391	72.4%	Mean: 81	FN: 48.8%, T: 51.2%	IF, A	1-year (28.1%)	OR
Bokshan (2018) [111]	United States	Retrospective	284	79.9%	Mean: 87.5	FN: 41.5%, T: 58.5%	-	1-year (-)	OR
Bottle (2006) [91]	England	Retrospective	129,522	79.4%	≥65	FN, T	IF: 48.0%, A: 40.7%, O: 11.3%	30-day (9.7%),	OR
Camur (2019) [146]	Turkey	Retrospective	109	65.1%	Mean: 80	FN: 62.4%, T: 37.6%	IF: 28.4%, HA: 71.6%	1-year (27.4%), 2-year (49.1%),	OR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
								5-year (80.2%)	
Camurcu (2017) [162]	Turkey	Retrospective	106	55.7%	Mean: 80.7	T: 100%	HA: 100%	1-year (34.0%)	OR
Cao (2021) [76]	Sweden	Retrospective	134,915	68.1%	Mean: 82.0	DFN: 37.2% , UFN: 13.6%, T: 47.3%, O: 1.9%	IF: 67.0%, HA: 25.7%, THA: 7.3%	30-day (7.6%)	OR
Carow (2017) [112]	Germany	Retrospective	437	74.8%	Mean: 81.2	T: 100%	IF: 100%	Inpatient (8.2%)	OR
Cenzer (2016) [113]	United States	Retrospective	857	76.0%	Mean: 83.8	FN: 100%	IF, O	1-year (27.4%)	OR
Chatterton (2015) [85]	England	Retrospective	4,426	74.5%	Mean: 82.0	FN: 74.0%, T: 26.0%	-	30-day (6.5%)	OR
Chen (2021) [147]	Taiwan	Prospective	281	70.1%	Mean: 81.3	FN: 52.7%, T: 47.3%	IF: 62.3%, HA: 37.7%	1-year (13.9%)	HR
Chiu (2018) [95]	Taiwan	Retrospective	6,626	64.4%	≥65	FN: 51.4%, T: 48.6%	IF: 60.4%, HA: 39.6%	30-day (1.6%), 3-month (3.6%), 1-year (10.2%)	HR
Crawford (2020) [104]	United States	Retrospective	5,918	-	≥60	DFN: 100%	HA: 100%	30-day (4.5%)	OR
D'Angelo (2005) [168]	Italy	Retrospective	299	84%	Mean: 80	DFN: 100%	HA: 100%	6-month (18.4%), 1-year (43.8%), 2-year (60.5%)	OR
De Luise (2008) [74]	Denmark	Retrospective	11,985	71.4%	Mean: 80	FN, T	IF, HA	30-day (-), 3-month (-), 1-year (-)	HR
Elliott (2003) [114]	Ireland	Prospective	1,780	76.7%	-	FN: 100%	-	1-year (22.0%)	OR
Endo (2005) [115]	United States	Retrospective	983	79.0%	Mean: 79.7	FN: 53.4%, T: 46.6%	-	1-year (11.0%)	HR
Eschbach (2013) [148]	Germany	Prospective	402	72.9%	≥60	FN: 48.5%, T: 51.5%	IF: 59.0%, A: 41.0%	Inpatient (6.2%)	OR
Faizi (2014) [92]	England	Retrospective	1,066	74.4%	Mean: 81	FN: 100%	-	30-day (8%)	OR
Fisher (2018) [169]	Australia	Retrospective	1,820	76.4%	Mean: 82.8	-	-	Inpatient (6.0%)	OR
Flodin (2016) [116]	Sweden	Prospective	843	73.2%	Mean: 82.1	FN: 51.6%, T: 48.4%	IF: 74.6%, A: 25.4%	1-year (15.2%)	OR
Folbert (2017) [117]	Netherlands	Retrospective	850	73.6%	Mean: 83	FN: 52.1%, T: 47.9%	IF: 65.9%, A: 33.6%,	1-year (23.2%)	OR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
							O: 0.5%		
Forni (2019) [77]	Italy	Prospective	728	77.6%	Mean: 83.8	FN, T	-	30-day (4.9%)	OR
Foss (2006) [86]	Denmark	Prospective	600	74.8%	Mean: 82.4	FN: 49.3%, T: 50.0%, O: 0.7%	-	30-day (13.5%)	OR
Franzo (2005) [97]	Italy	Retrospective	6,629	81.3%	Mean: 82.4	FN, O	IF, HA, THA	Inpatient (5.4%), 30-day (9.6%), 6-month (20.0%), 1-year (25.3%)	OR
Fu (2021) [149]	China	Retrospective	528	75.5%	Mean: 77.9	FN: 100%	HA, THA	1-year (23.4%)	OR
Giummarra (2020) [118]	Australia	Retrospective	4,621	69.5%	Mean: 83.4	FN, T	IF: 54.6%, HA: 35.7%, THA: 3.9%, O: 5.8%	1-year (29.4%)	HR
Härstedt (2015) [164]	Sweden	Prospective	272	72.1%	Mean: 82.6	-	-	6-month (13.2%)	OR
Henderson (2015) [119]	Ireland	Prospective	206	73%	Median: 82	FN: 100%	-	1-year (12.1%)	OR
Heyes (2017) [157]	Ireland	Retrospective	443	70.2%	Mean: 77	-	IF, HA, THA	1-year (15.1%)	OR
Ho (2010) [120]	Taiwan	Retrospective	409	51.6%	Mean: 72.5	FN: 40.6%, T: 59.4%	IF: 73.3%, A: 26.7%	1-year (14%)	HR
Huette (2020) [121]	France	Prospective	309	73.5%	Median: 85	-	IF: 47.2%, HA: 37.5%, THA: 12.6%, O: 2.7%	1-year (23.9%)	HR
Hung (2017) [122]	Taiwan	Retrospective	5,982	64.8%	Mean: 74.9	FN: 45.2%, T: 45.6% O: 9.2%	IF: 57.8%, A: 42.2%	3-month (-), 1-year (-), 6-year (-), 10-year (34.2%)	HR
Ireland (2015) [100]	Australia	Retrospective	2,552	62.4%	Mean: 86.6	FN: 38.4%, T: 46.7%, O: 14.9%	IF: 38.4%, HA: 42.4%, THA: 4.3%, O: 14.9%	30-day (11%), 1-year (34%), 2-year (47%)	HR
Ishidou (2017) [155]	Japan	Prospective	377	81.4%	Mean: 83.1	FN: 36.4%, T: 63.6%	IF: 70.3%, HA: 26.8%,	1-year (8.0%)	OR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
							O: 2.9%		
Jiang (2005) [123]	Canada	Retrospective	3,981	71.3%	Median: 82	-	-	Inpatient (6.3%)	OR
Kang (2010) [124]	Korea	Retrospective	9,817	70.2%	Mean: 74.9	FN: 55.4%, T: 38.0%, O: 6.6%	IF: 46.8%, HA: 2.2%, THA: 46.6%, O: 4.4%	1-year (16.6%)	HR
Kannegaard (2010) [156]	Denmark	Retrospective	42,076	73.1%	Mean: 80.7	FN: 58.7%, T: 36.8%, O: 4.5%	-	1-year (29.3%)	HR
Karres (2018) [90]	Netherlands	Retrospective	746	57.1%	Mean: 80	FN: 44.5%, T: 55.5%	IF, HA	30-day (8.2%)	OR
Khan (2013) [105]	England	Retrospective	467	72.6%	Mean: 79.6	FN: 55.5%, T: 44.5%	-	30-day (7.5%)	OR
Kim (2012) [125]	Korea	Prospective	415	68.2%	Mean: 75.1	FN: 55.7%, T: 44.3%	IF: 43.9%, A: 56.1%	1-year (14.7%)	HR
Kim (2016) [160]	Korea	Retrospective	772	75.1%	Mean: 79.4	FN: 46.2%, T: 53.8%	-	1-year (14.1%)	OR
Kirkland (2011) [99]	United States	Retrospective	485	73.4%	Mean: 82.3	-	-	30-day (8.2%)	OR
Kovar (2015) [171]	Austria	Retrospective	3,595	72.2%	Mean: 78.5	FN: 43.8%, T: 53.2%, O: 3.0%	-	3-month (10.7%), 6-month (11.4%), 1-year (12.2%)	OR
Lau (2015) [166]	China	Retrospective	759	72%	Mean: 84	DFN: 25%, UFN: 24%, T: 51%	IF: 75%, HA: 25%	30-day (2.5%), 1-year (16.3%)	HR
Lawrence (2017) [175]	England	Retrospective	1,979	70.6%	Median: 84	DFN: 48.8%, UFN: 9.0%, T: 42.0%, O: 0.2%	IF: 50.2%, HA: 46.4%, THA: 3.4%	1-year (24.4%)	HR
Lizaur-Utrilla (2016) [158]	Spain	Prospective	628	74.2%	Mean: 83.5	FN: 37.1%, T: 62.9%	IF, A	Inpatient (0.9%), 3-month (7.0%), 1-year (13.6%)	OR
Lizaur-Utrilla (2019) [94]	Spain	Prospective	1,083	71.4%	Mean: 83.3	FN: 31.0%, T: 69.0%	-	Inpatient (8.2%)	OR
Mangoni (2013) [150]	Netherlands	Retrospective	71	70.4%	Mean: 84	FN: 48.5%, T: 42.5%, O: 8.8%	IF: 61.8%, A: 38.2%	3-month (12.7%), 1-year (25.4%)	HR
Mariconda (2015) [161]	Italy	Prospective	552	77.3%	Mean: 78.3	FN: 43.7%, T: 56.3%	IF: 60.2%, HA: 27.6%,	30-day (4.3%),	HR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
							THA: 12.2%	1-year (18.8%)	
Maxwell (2008) [93]	England	Prospective	4,967	76.4%	Mean: 79.9	FN: 100%	-	30-day (8.0%)	OR
Mayordomo-Cava (2020) [78]	Spain	Retrospective	5,543	79.1%	Mean: 93.2	T: 57.7%, O: 42.3%	IF: 60.9%, O: 39.1%	30-day (7.0%)	OR
Mazzola (2015) [126]	Italy	Prospective	275	85.5%	Mean: 89.4	FN: 44.4% T: 55.6%	IF, HA	6-month (21.2%)	OR
Menéndez-Colino (2018) [127]	Spain	Prospective	491	79.2%	Mean: 85.6	T: 58.0%, O: 42.0%	IF: 54.6%, A: 37.5%, O: 7.9%	1-year (23.2%)	HR
Meng (2021) [128]	China	Retrospective	480	66.5%	Mean: 78.3	FN, T	IF: 61.3%, HA: 26.7%, THA: 12.0%	1-year (15.6%)	OR
Morrissey (2017) [79]	England	Retrospective	1,913	73.7%	Mean: 83.9	FN: 49.7%, T: 50.3%	IF: 47.5%, HA: 44.5%, THA: 6.2%, O: 1.8%	30-day (6.1%)	OR
Myers (1991) [130]	United States	Retrospective	27,370	80.0%	Median: 81	FN: 29.8%, T: 52.8%, O: 17.4%	IF, HA, THA, O	Inpatient (4.9%)	OR
Nijland (2017) [80]	Netherlands	Retrospective	1,803	67.3%	Median: 83	FN: 45.1%, T: 54.0% O: 0.9%	IF: 65.6%, HA: 34.4%	30-day (7.6%), 1-year (29.0%)	OR
Nijmeijer (2016) [87]	Netherlands	Retrospective	850	73.6%	Median: 83.0	FN: 52.1%, T: 41.9%	-	30-day (7.5%)	OR
Norring-Agerskov (2017) [173]	Denmark	Retrospective	7,293	76.6%	Mean: 82.6	-	-	30-day (10.2%)	HR
Norring-Agerskov (2019) [89]	Denmark	Retrospective	113,211	68.8%	Median: 81	FN, T	-	30-day (10.1%)	OR
Nuotio (2016) [131]	Finland	Prospective	472	75.2%	Median: 82	-	-	4-month (19.1%)	OR
O'Daly (2010) [132]	Ireland	Retrospective	377	78.0%	Median: 83	FN: 50.9%, T: 49.1%	IF, HA	1-year (25.5%)	HR
Padrón-Monedero (2017) [133]	Spain	Retrospective	31,884	75.9%	$\geq$65	-	IF, A	Inpatient (5.5%)	OR
Pang (2020) [81]	England	Retrospective	894	72.8%	Mean: 82.7	FN: 100%	-	30-day (9.5%)	OR
Pereira (2010) [134]	Brazil	Prospective	246	72.8%	Mean: 79.3	-	IF: 77.6%, A: 22.4%	1-year (35%)	HR
Petersen (2006) [135]	Denmark	Retrospective	1,186	77.3%	Mean: 81.6	DFN: 100%	HA: 100%	3-month (-), 1-year (-)	HR
Petersen (2020) [96]	Denmark	Retrospective	11,318	71.6%	Median: 84	FN: 51.7%, T: 46.1%,	IF: 63.7%, A: 36.3%	30-day (11.4%)	HR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
						O: 2.2%			
Pioli (2006) [159]	Italy	Prospective	248	85.5%	Mean 83.6	FN: 46.0%, T: 54.0%	-	Inpatient (4.8%), 3-month (12.9%), 6-month (19.0%), 1-year (25.4%)	OR
Rae (2007) [84]	Australia	Prospective	222	72.1%	Mean: 79	FN: 100%	IF: 63.0%, HA: 36.0%, THA: 1.0%	30-day (7.2%)	OR
Ribeiro (2014) [136]	Brazil	Prospective	418	76.1%	Mean: 79.8	FN: 38.8%, T: 61.2%	IF: 53.4%, HA: 30.6%, THA: 6.0%, O: 10.0%	Inpatient (4.3%), 1-year (15.3%)	Both
Roche (2005) [101]	England	Prospective	2,448	79.9%	Mean: 82	FN: 57%, T: 43%	-	30-day (9.6%), 1-year (33%)	HR
Rosso (2016) [137]	Italy	Retrospective	1,448	75.8%	Mean: 80.3	-	-	30-day (4.0%), 6-month (14.1%), 1-year (18.8%)	OR
Sanz-Reig (2018) [138]	Spain	Prospective	331	73.1%	Mean: 83.7	FN: 57.7%, T: 42.3%	IF: 40.2% HA: 53.5% O: 6.3%	Inpatient (11.4%)	OR
Schuijt (2021) [88]	Netherlands	Retrospective	492	56.9%	Median: 89	-	-	30-day (12.4%)	OR
Sheikh (2017) [102]	England	Retrospective	1,356	72.8%	Mean: 81.4	DFN: 31.0%, UFN: 31.2%, T: 32.1%, O: 5.7%	IF, HA	30-day (8.7%)	HR
Söderqvist (2006)	Sweden	Prospective	213	81%	Mean: 84	T: 100%	IF: 100%	1-year (24.9%)	HR
Söderqvist (2009) [139]	Sweden	Prospective	1,944	74.4%	Mean: 84	UFN: 13.2%, DFN: 36.4%, T: 50.4%	-	4-month (16%), 2-year (38%)	HR
Tal (2016) [140]	Israel	Retrospective	1,161	73.9%	Mean: 81.8	FN: 42.4%, T: 57.6%	IF, HA	Inpatient (8.8%), 1-year (20.6%)	OR
Talsnes (2011) [141]	Norway	Prospective	302	76.2%	Mean: 84.5	DFN: 100%	HA: 100%	3-month (19.5%)	OR

Continued on next page

Table A.3 (Continued)

Study (first author and year)	Country	Study design	Sample	Females	Population age	Fracture types	Surgical procedures	Mortality (rate)	Risk
Thomas (2014) [82]	England	Retrospective	2,989	71.7%	Mean: 81	FN: 100%	-	30-day (8.8%)	OR
Thorne (2021) [163]	England	Retrospective	2,422	70.6%	Median: 85	FN: 100%	IF: 40.4%, A: 47.2%, O: 12.4%	Inpatient (6.4%), 1-year (23.5%)	OR
Van de Ree (2020) [83]	Netherlands	Retrospective	925	69.9%	Mean: 81.9	FN, T	IF, HA, THA	30-day (9.9%)	OR
Velez (2020) [170]	Colombia	Retrospective	275	70.5%	Mean: 79.9	T: 100%	IF: 100%	6-month (16.0%)	OR
Vosoughi (2017) [142]	Iran	Prospective	724	56.1%	Mean: 75.7	FN: 24.2%, T: 75.8%	IF: 91.6%, A: 8.4%	3-month (14.5%), 1-year (22.4%)	OR
Wang (2021) [103]	China	Retrospective	460	67.0%	Mean: 79.3	FN: 52.4%, T: 47.6%	-	30-day (4.1%), 6-month (13.3%), 1-year (20.0%)	HR
Wu (2016) [167]	Taiwan	Prospective	195	48.2%	Mean: 79.4	FN: 63.6%, T: 46.4%	IF: 46.4%, HA: 33.8%, THA: 20.0%, O: 12.2%	1-year (20.0%)	HR
Würdemann (2021) [75]	Netherlands	Prospective	4,421	67.6%	Mean: 79.9	DFN: 41.7%, UFN: 14.8%, T: 43.5%	-	30-day (7.2%), 3-month (12.5%), 1-year (21.1%)	OR
Xing (2021) [143]	China	Retrospective	445	61.6%	Mean: 79.4	FN: 55.7, T: 44.3%	IF: 51.5%, A: 48.5%	1-year (14.4%)	OR
Yombi (2019) [144]	Belgium	Retrospective	829	67.8%	Mean: 81	-	IF: 64.3%, HA: 22.3%, THA: 11.8%, O: 1.6%	1-year (23.5%)	HR
Yoo (2018) [152]	Korea	Retrospective	324	75.9%	Mean: 77.8	FN: 38.6%, T: 64.8%	IF: 33.0%, A: 64.8%, O: 2.2%	1-year (9.0%)	HR
Zanetti (2019) [165]	Italy	Retrospective	1,211	78.6%	Mean: 84.7	-	A: 57.0%, O: 43%	3-month (11.4%), 6-month (17.0%), 1-year (23.5%)	OR

A.2 Risk of Bias Assessment Protocol

Study Participation

Prompts

1. Were the selection criteria clearly defined?
    - Yes: Inclusion and exclusion criteria were clearly reported.
    - No: Solely a description of study participants (e.g. number and types of participants described, but no further details on the selection process).
2. Was the study population representative of the population of interest?
    - Yes: Both intracapsular AND extracapsular fractures were present in the study population, without exclusion based on residence OR cognitive status OR mobility status[1], AND at most 10% of sample was unrepresentative (pathological fractures, periprosthetic fractures, conservative treatments). If the paper does not explicitly report on intracapsular and extracapsular fractures, but instead broadly mentions that "all hip fracture patients" were enrolled into the study, it is assumed that both intracapsular and extracapsular fractures are represented.[2]
    - No: Evidence was found that one of the aforementioned conditions had not been satisfied.
3. Were the baseline characteristics of the study sample described in sufficient detail?
    - Yes: Baseline characteristics were described in summary tables using descriptive statistics (e.g. gender and age distributions, prevalence of comorbidities, fracture types).
    - No: The study sample was only described in a general manner, without descriptive statistics (e.g. the sample comprised older adults above the age of 70 with trochanteric and femoral neck fractures).

Rating scheme

- Low risk of bias if:
    - All questions were answered with yes.
- Moderate risk of bias if:
    - At most one question was answered with no.
- High risk of bias if:

[1] The criteria of not excluding patients based on preoperative residence, cognitive status, and mobility status was defined after amendment of the protocol. The first and second author added these specifications to make the notion of a representative population less ambiguous.

[2] The last assumption was added after amendment of the protocol. Not all papers specifically reported on the exact types of fractures in their cohorts. It was deemed unlikely that descriptions along the line of "all hip fracture patients" would exclude either intracapsular or extracapsular fractures.

  - At least two questions were answered with no.

Attrition

Prompts

1. Was completeness of follow-up adequate?
   - Yes: At most 20% of the data was lost.
   - No: Data loss above the pre-defined threshold.
   - Unclear: No statement on loss of follow-up was provided.
2. Were there no important differences between participants who completed the study and those who did not?
   - Yes: The data appeared to be missing at random.
   - No: Missing data was correlated with relevant patient characteristics.
   - Unclear: No characteristics of patients lost to follow-up were provided OR no conclusion could be drawn since no statement for data loss was provided in the first place.

Rating scheme

- Low risk of bias if:
  - All questions were answered with yes.
- Moderate risk of bias if:
  - Both questions were answered with unclear.
  - Only question 1 was answered with no.
- High risk of bias if:
  - Question 2 was answered with no.

Prognostic Factor Measurement

Prompts

1. Were prognostic factors clearly defined?
   - Yes: At least 80% of the prognostic factors (i.e. covariates in multiple regression models) were defined using clear medical classifications (e.g. ICD codes, or clear cut-off levels where applicable). For sum scores, e.g. the Nottingham Hip Fracture Score or the Charlson Comorbidity Index, it suffices to provide definitions of proper cut-off levels. That is, the exact descriptions of the comorbidities contributing to the sum scores do not have to be provided.
   - No: Less than 80% of the prognostic factors were not described in sufficient detail to permit replication.

2. Were prognostic factors measured or acquired appropriately?
    - Yes: Prognostic factors were measured preoperatively AND obtained through routinely used medical practices (e.g. chart abstraction or chart review).
    - No: No clear indication that prognostic factors were present preoperatively OR prognostic factors were defined based on self-report.

Rating scheme

- Low risk of bias if:
    - All questions were answered with yes.
- Moderate risk of bias if:
    - At most one question was answered with no.
- High risk of bias if:
    - Both questions were answered with no.

Outcome Measurement

1. Is the outcome measure clearly and correctly defined?
    - Yes: Inpatient (including days of admission), 30-day, or 1-year mortality are reported.
    - No: In case of inpatient mortality, the length of admission is left unspecified.
2. Has the outcome measurement been obtained validly?
    - Yes: In prospective studies, mortality was obtained through follow-up interviews with patients and relatives. In retrospective studies, mortality was obtained from medical records or municipal mortality registers. Furthermore, 30-day mortality should be measured including discharged patients, and not solely inpatients.
    - No: No description of outcome ascertainment is provided OR 30-day mortality is solely based on inpatient records.

Rating scheme

- Low risk of bias if:
    - All questions were answered with yes.
- Moderate risk of bias if:
    - At most one question was answered with no.
- High risk of bias if:
    - Both questions were answered with no.

Confounding

Prompts and rating

1. Were important confounders accounted for?
    - Yes: Low risk if regression analyses controlled for at least age AND gender, AND (comorbidities (e.g. >2, ASA, CCI), OR preoperative functional status in ADL, OR preoperative mobility OR living situation OR cognitive status). Sum scores which include age, gender and comorbidities, such as the Nottingham Hip Fracture Score are assumed to provide sufficient control for confounding as well[3].
    - Yes: Moderate risk if regression analyses controlled for at least age AND gender.
    - Yes: Moderate risk if regression analyses controlled for either age OR gender AND one of the other confounders listed above.
    - No: High risk if neither of the aforementioned combinations of confounders was controlled for.

Analysis and Reporting

1. Is the statistical model appropriate for the study design?
    - Yes: Appropriate regression models were used to analyse the influence of prognostic factors in cohort studies (e.g. logistic regression or Cox proportional hazard regression).
    - No: None of the aforementioned regression models were used in the statistical analysis.
2. Were results reported without selective biases?
    - Yes: Risk measures of all covariates included in the multivariate regression models were reported.
    - No: Risk measures of covariates included in multivariate models were suppressed, based on statistical insignificance.

Rating scheme

- Low risk of bias if:
    - All questions were answered with yes.
- Moderate risk of bias if:
    - At most one question was answered with no.
- High risk of bias if:

[3]The protocol for dealing with sum scores was an amendment following an initial trial phase of the risk of bias assessment. Initially, only the individual confounding factors were considered.

– Both questions were answered with no.

Overall Risk of Bias Assessment

1. The overall risk of bias of a study was rated as low if:
   - All domains are bias domains were rated low risk.
   - The attrition domain was rated to be at moderate risk of bias AND all other domains were rated to be at low risk of bias.[4]
2. The overall risk of bias of a study is rated as moderate if:
   - At least two domains are rated moderate risk, and all other domains are rated low risk.
3. The overall risk of bias of a study is rated high if:
   - At least one domains was rated to be at high risk of bias.
   - At least four domains were rated to be at moderate risk of bias.

A.3 Risk of Bias Summary

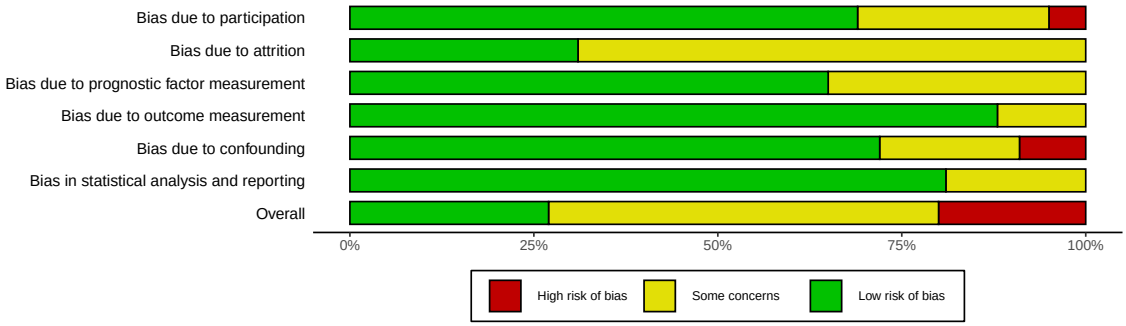

Figure A.1: Unweighted risk of bias summary of all studies included in the systematic review examining predictors for mortality within one year following hip fracture surgery.

[4] A priori, it had been decided that moderate risk of bias was acceptable for an overall low risk of bias. It was anticipated that data missing to follow-up would not be reported in detail in many cases. However, it was expected that if data were lost to follow-up, that this would mostly concern small percentages with limited influence on the inferences, as retrospective cohort designs generally allow for the analysis of very large sample sizes.

A.4 Certainty of Evidence Assessment (GRADE)

Risk of Bias

Rate down by one level if the cumulative weight of studies at high risk of bias exceeds 60% in an analysis.

Inconsistency

Rate down by one level if I^2 exceeds 60%, with unexplained causes for heterogeneity. Exceptions apply to cases where the interpretation of I^2 is misleading due to presence of large studies which report exceptionally narrow confidence intervals. In these cases, the forest plots should be examined further to determine whether the majority of the confidence intervals overlap, and whether the point estimates are similar. If this is the case, decide against downgrading for inconsistency.

Imprecision

Rate down by one level if both the Knapp-Hartung 95% confidence interval (CI) and the Bayesian 95% credible interval (CrI) overlap with the null effect.

Rate down by two levels if the DerSimonian-Laird, Knapp-Hartung, and Bayesian C(r)Is all overlap with the null effect.

Publication Bias

Rate down by one level if the corrected pooled estimate produced by either the L_0^+ estimator OR the R_0^+ estimator of the trim-and-fill method is 20% lower than the DerSimonian-Laird estimate. Solely the suspicion of studies being suppressed due to publication bias is not sufficient for downgrading: the magnitude of the effect of publication bias is of interest.

A.5 Bayesian Hierarchical Model Specification

This supplement discusses the Bayesian hierarchical model which was used for the Bayesian sensitivity analyses with respect to heterogeneity underestimation. The Bayesian hierarchical model was specified as follows, with parameters based on the model by Harrer et al.[5] (A.1).

$$\begin{cases} \hat{\theta}_i \sim \mathcal{N}\left(\theta_i, \sigma_i^2\right) \\ \theta_i \sim \mathcal{N}(\mu, \tau^2) \\ \mu \sim \mathcal{N}(0, 1) \\ \tau \sim \mathcal{HC}(0, 0.5) \end{cases} \tag{A.1}$$

where $\hat{\theta}_i$ denotes the observed effect size in study i, which deviates from the true underlying effect θ_i due to sampling errors which are captured by the variance term σ_i^2. It is assumed that θ_i is drawn from a normal distribution with mean μ and variance τ^2. The latter two parameters require manual specification by researchers. The underlying mechanics and rationales are briefly described here, but the reader is referred to the book by Harrer et al.[5] for additional information.

A common problem in meta-analyses involving only a few studies is that the between-study variance τ^2 is often estimated at exactly zero, even though this is highly unlikely to be correct[6]. The estimated effects of prognostic factors are bound to differ due to population and intervention heterogeneity. Failing to acknowledge between-study heterogeneity leads to overly confident pooled estimates and potentially false positives in significance testing.

The Bayesian hierarchical model allows prior knowledge and assumptions to be incorporated into the pooling procedure through so-called weakly informative priors. Weakly informative priors gently influence the analytical results based on researchers' theoretically supported assumptions, without biasing the results too strongly towards researchers' beliefs. One such valid assumption is that τ^2 should never be zero. In the proposed Bayesian hierarchical model, this is ensured by modelling the between-study variance with a Half-Cauchy (HC) prior, which is one of the most used weakly informative priors for modelling variance terms in Bayesian statistics[7]. This modelling step is represented by $\tau^2 \sim \mathcal{HC}(0, 0.5)$ in the hierarchical model, which realistically represents values for τ encountered in meta-analyses[5].

To complete the modelling process of θ_i, only μ is left to be specified. This is the value that the true underlying effect size attains on average. Since this is challenging to anticipate, we limit the assumptions imposed by the prior. Choosing $\mu \sim \mathcal{N}(0, 1)$ entails a 95% chance that the mean log odds or mean log hazards attains a value between -2 and 2. This is a conservative guess, since the prior spans both positive and negative ranges: no assumptions are made about the direction of the prognostic factors' effects.

[5] Harrer M, Cuijpers P, A FT et al. Doing Meta-Analysis With R: A Hands-On Guide. 1st ed. Boca Raton, FL and London: Chapman & Hall/CRC Press, 2021

[6] Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med 2013;32:4071-89.

[7] Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 2006; 1:515-34.

A.6 Forest Plots with Risk of Bias Assessments

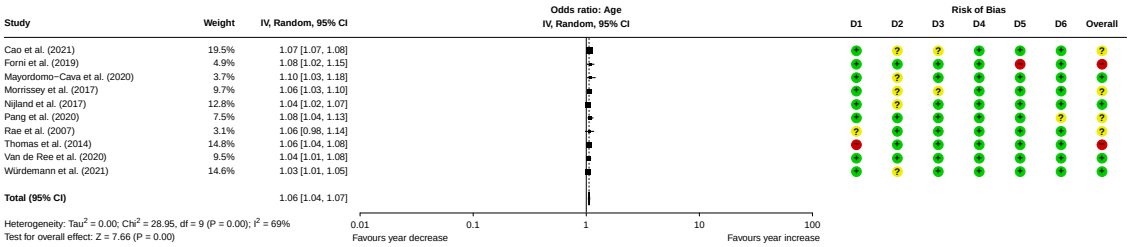

Figure A.2: Forest plot illustrating the increased odds of 30-day mortality for advanced age (per year increase). High-quality evidence was found for age being a risk factor for 30-day mortality following hip fracture surgery.

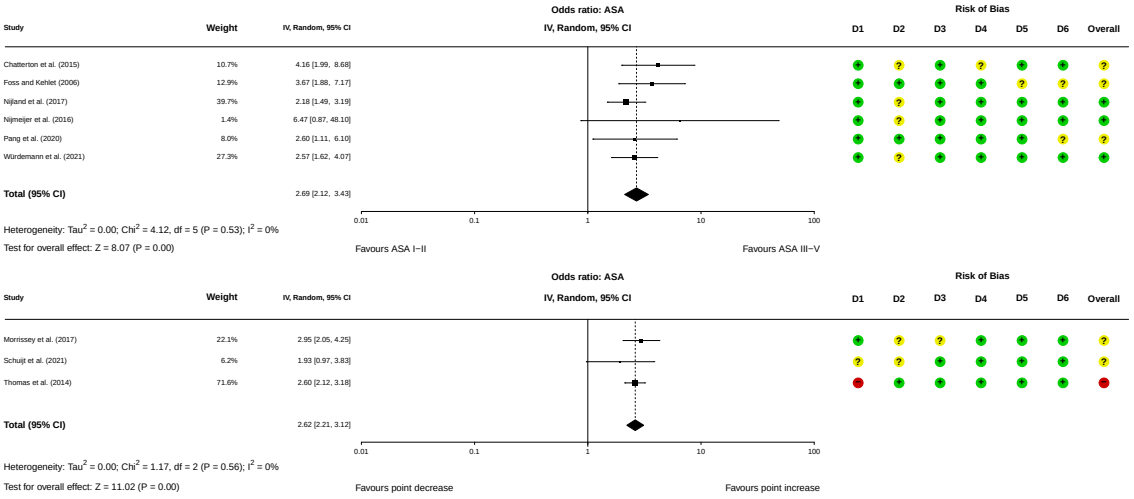

Figure A.3: Forest plots illustrating the increased odds of 30-day mortality for increased ASA scores. The upper panel shows the influence of ASA III-V compared to ASA I-II. The lower panel shows the influence per point increase. Based on the pooled estimate comprising most data (upper panel), ASA scores were rated as a high-quality evidence predictor for 30-day mortality following hip fracture surgery.

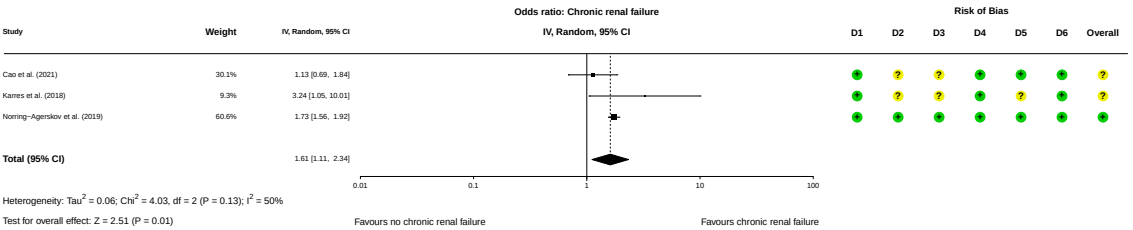

Figure A.4: Forest plot illustrating the increased odds of 30-day mortality in presence of comorbid chronic renal failures. Moderate-quality evidence was found for chronic renal failure.

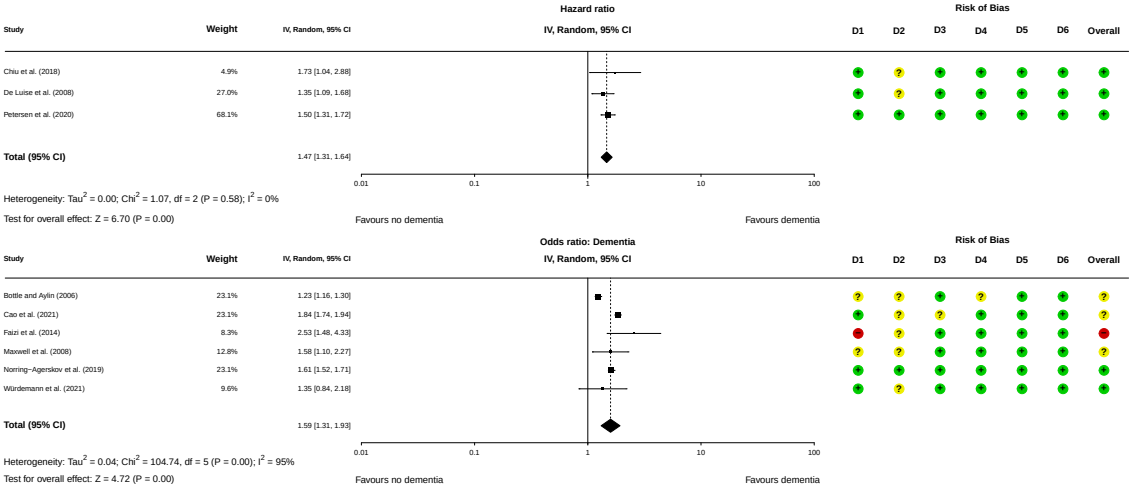

Figure A.5: Forest plots illustrating the increased hazards and odds of 30-day mortality in presence of comorbid dementia. Based on the pooled estimate comprising most data (lower panel), dementia was rated as a moderate-quality evidence predictor for 30-day mortality following hip fracture surgery.

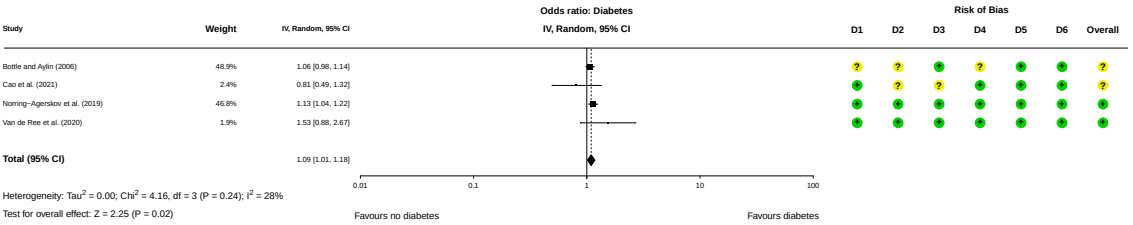

Figure A.6: Forest plot illustrating the increased odds of 30-day mortality in presence of comorbid diabetes. Moderate-quality evidence was found for diabetes.

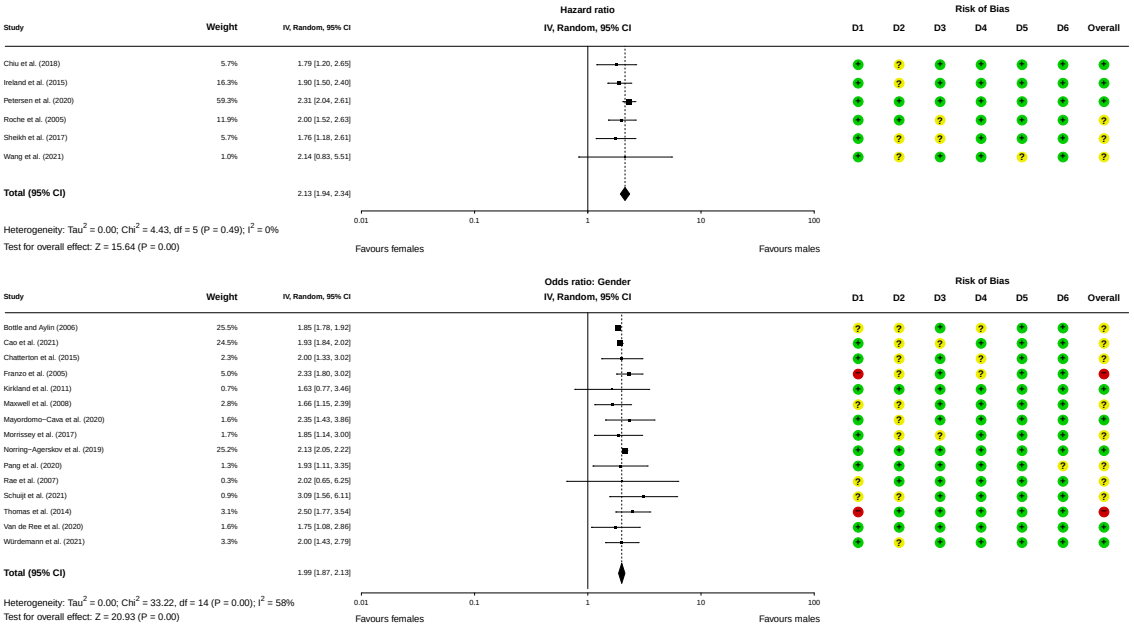

Figure A.7: Forest plots illustrating the increased hazards and odds of 30-day mortality amongst males compared to females. Both pooled estimates were found to be of high-quality evidence.

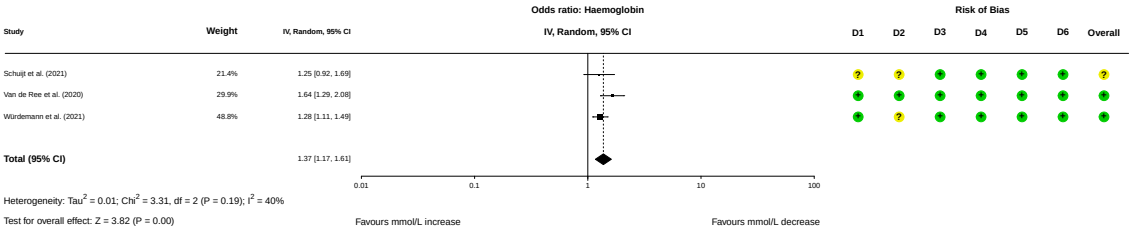

Figure A.8: Forest plot illustrating the increased risk of 30-day mortality in case of lower haemoglobin levels at admission. Moderate-quality evidence was found for low haemoglobin.

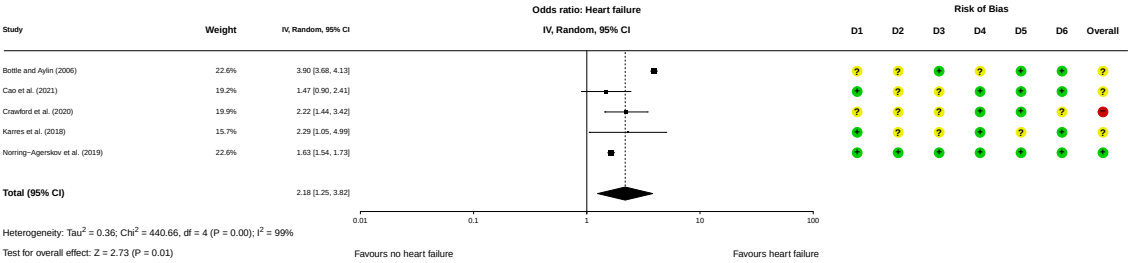

Figure A.9: Forest plot illustrating the increased risk of 30-day mortality in presence of comorbid heart failures. Moderate-quality evidence was found for heart failure.

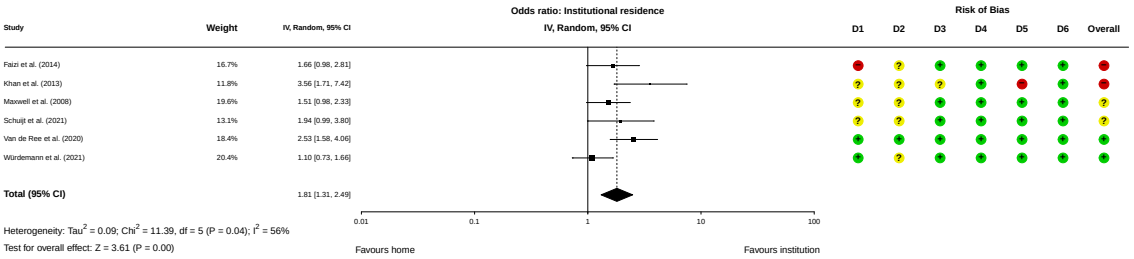

Figure A.10: Forest plot illustrating the increased risk of 30-day mortality for patients living in an institution prior to admission. High-quality evidence was found for institutional.

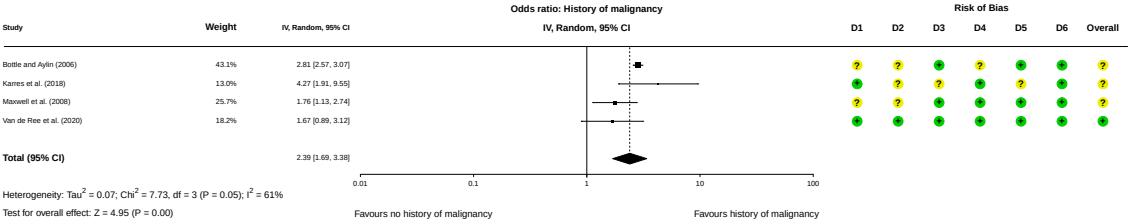

Figure A.11: Forest plots illustrating the increased risk of 30-day mortality for patients with a history of any malignancy. Moderate-quality evidence was found for a history of any malignancy.

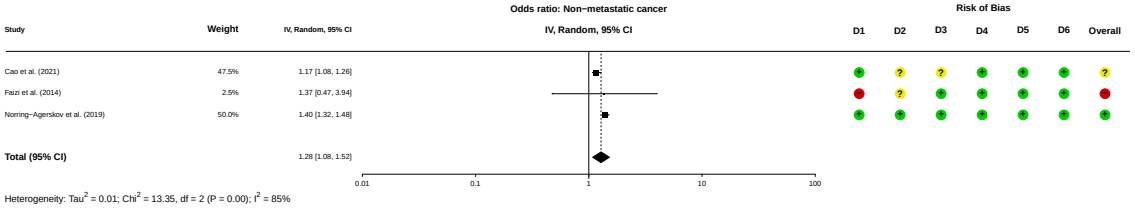

Figure A.12: Forest plots illustrating the increased risk of 30-day mortality for patients with non-metastatic cancer. Low-quality evidence was found for non-metastatic cancer.

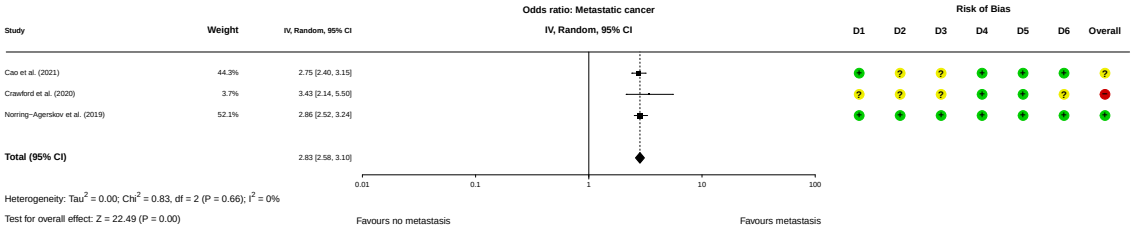

Figure A.13: Forest plots illustrating the increased risk of 30-day mortality for patients with metastatic cancer. High-quality evidence was found for non-metastatic cancer.

APPENDIX B

Vignette Study Supplements

B.1 D-efficiency of Experimental Designs

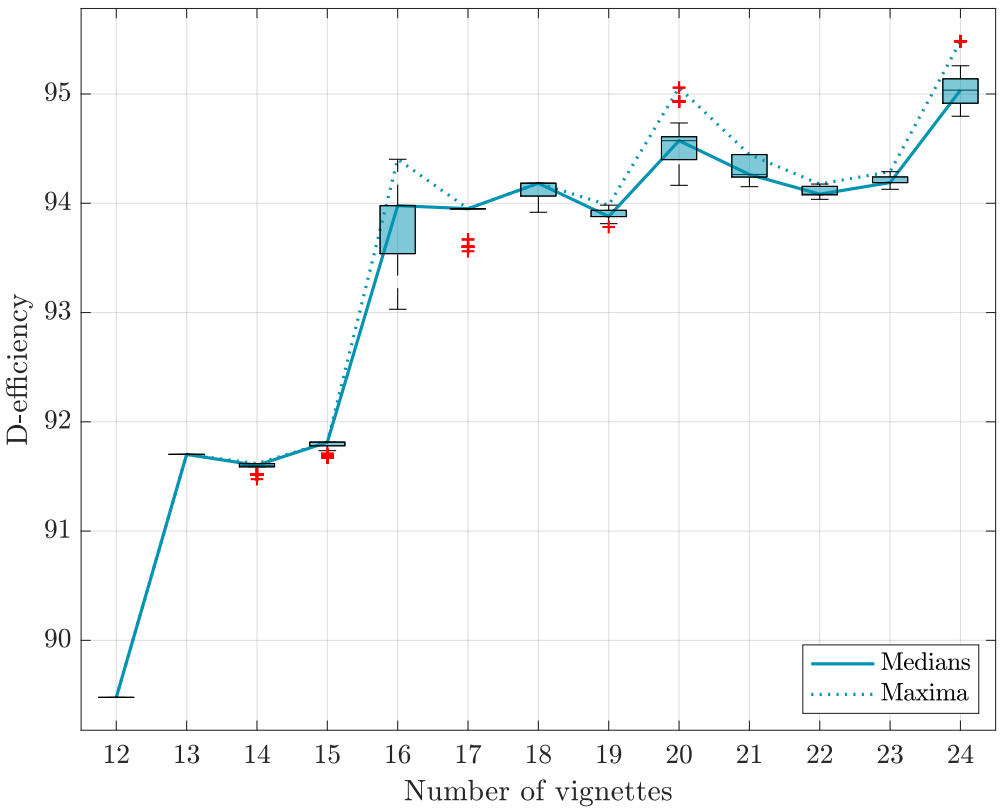

Figure B.1: Distribution of D-efficiencies for a fixed number of vignettes, each computed for 300 different initial conditions in Federov's exchange algorithm. Outliers are marked in red.

B.2 Calibration Questions

Table B.1: Overview of calibration questions and their true seed realisations, accompanied by their 95% confidence intervals (CIs) computed through Rubin's rules.

Calibration Question	Realisation (95% CI)
How many percent of the **female** hip fracture patients aged **80 years or older** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	8.0% (4.8-11.3%)
How many percent of the **male** hip fracture patients aged **90 years or older** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	18.3% (13.4-23.1%)
How many percent of the hip fracture patients aged **85 years or older** with an **ASA IV classification** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	25.9% (20.1-31.7%)
How many percent of the hip fracture patients aged **80 years or older** with an **ASA II-III classification** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	8.2% (4.9-11.5%)
How many percent of the hip fracture patients aged **80 years or older** with a **high risk of malnutrition** (SNAQ score ≥ 3) and **pre-fracture institutional residence** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	22.8% (16.6-29.0%)
How many percent of the hip fracture patients aged **80 years or older** with a **high risk of malnutrition** (SNAQ score ≥ 3) and **preoperative anaemia** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	13.9% (9.2-18.7%)
How many percent of the hip fracture patients aged **80 years or older** with a **displaced femoral neck fracture** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	6.5% (3.4-9.7%)
How many percent of the hip fracture patients aged **80 years or older** who were **fully independent in activities of daily living** (Katz score of 6) and at **low risk of malnutrition** (SNAQ score ≤ 1), died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	4.2% (1.7-6.6%)
How many percent of the hip fracture patients aged **90 years or older**, who were **mobile without walking aids** and **did not have dementia**, died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	10.2% (6.3-14.2%)
How many percent of the hip fracture patients aged **80 years or older** with an **ASA IV classification** and **prefracture institutional residence** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	33.2% (26.7-39.7%)
How many percent of the hip fracture patients aged **90 years or older** with an **extracapsular fracture** and **preoperative anaemia** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	15.7% (11.1-20.3%)
How many percent of the hip fracture patients aged **90 years or older** with an **ASA I-II classification** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	3.9% (1.4-6.4%)
How many percent of the hip fracture patients aged **90 years or older** with an **ASA III-IV classification**, **dementia** and **pre-fracture institutional residence** died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	23.2% (17.6-28.7%)
How many percent of the hip fracture patients aged **90 years or older** with **severe functional handicaps** (Katz score 0-2), died within 30 days following hip fracture surgery between 2017-2019, according to the DHFA-TFI group?	16.7% (12.1-21.3%)

B.3 MCMC Convergence Diagnostics

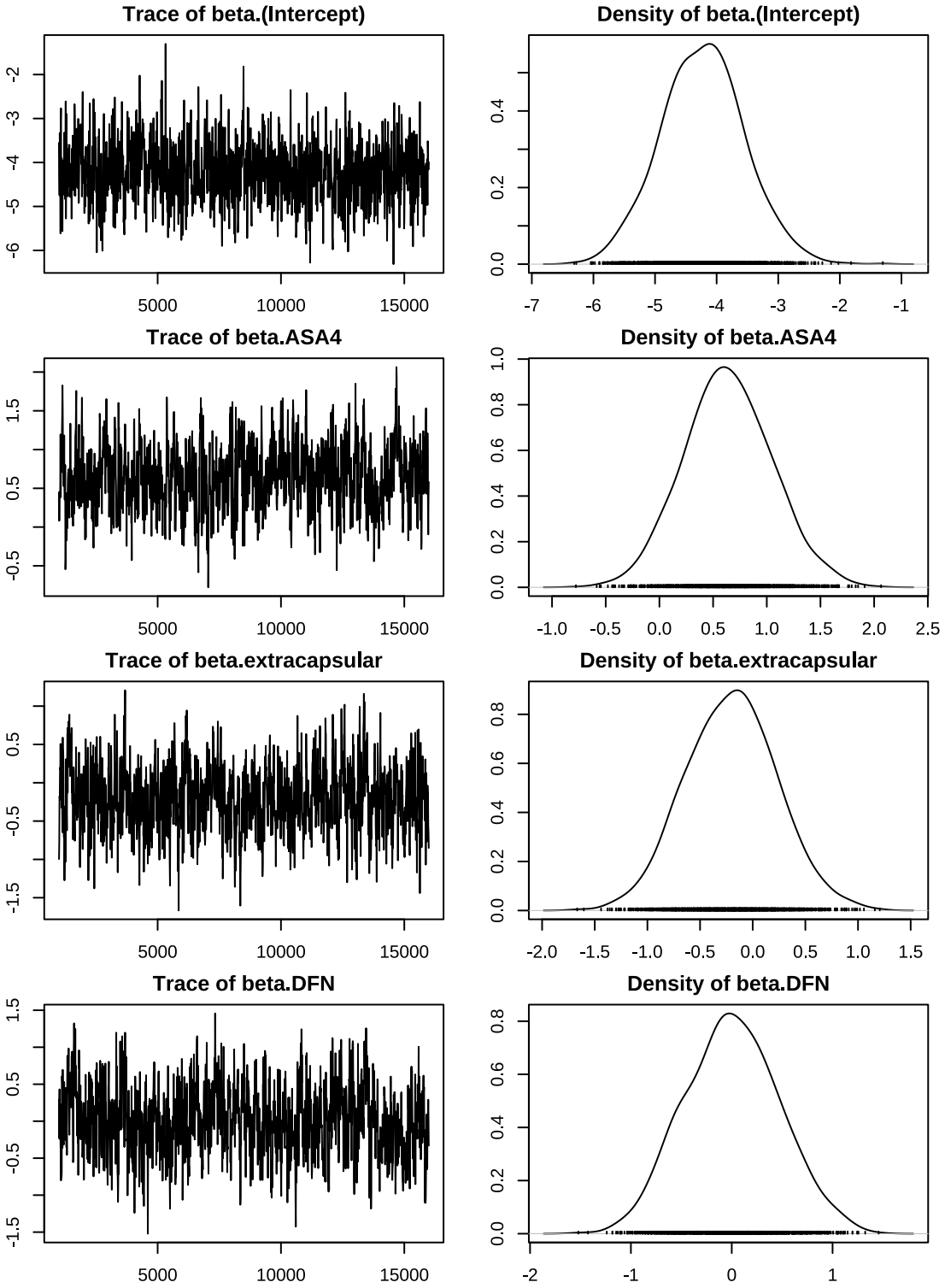

Figure B.2: Trace plots and densities of the posterior distributions of the β-coefficients (1/3).

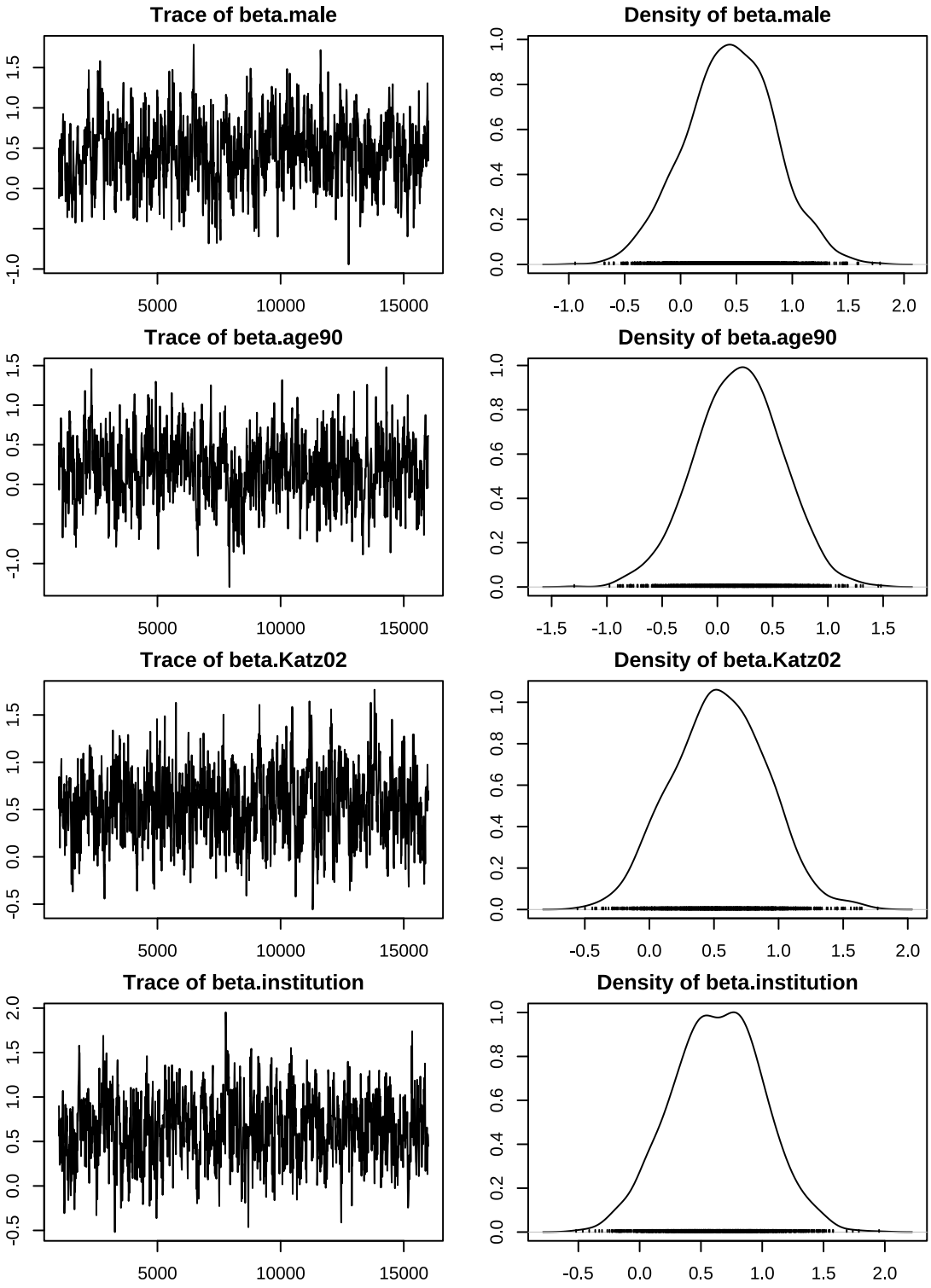

Figure B.3: Trace plots and densities of the posterior distributions of the β-coefficients (2/3).

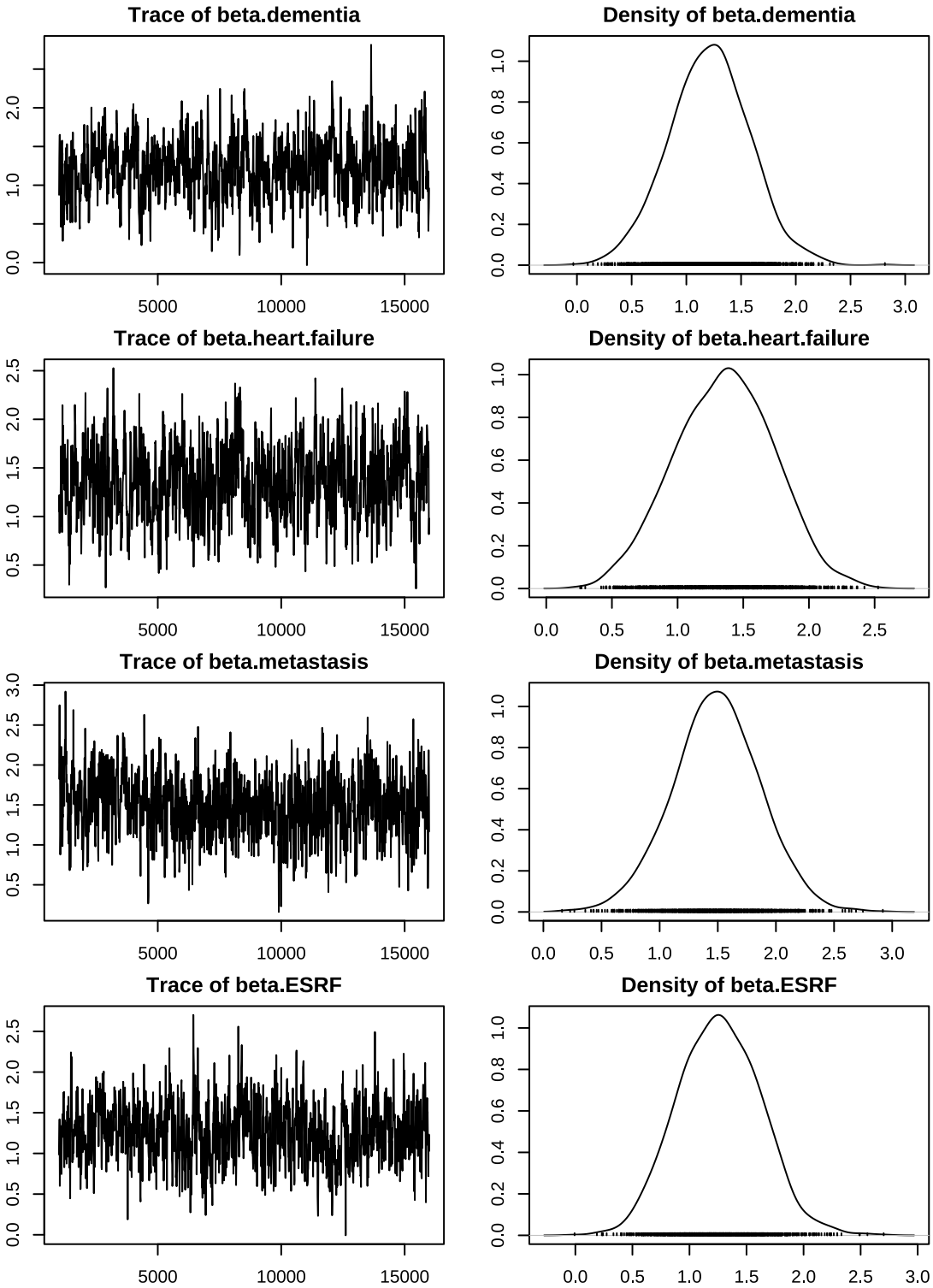

Figure B.4: Trace plots and densities of the posterior distributions of the β-coefficients (3/3).

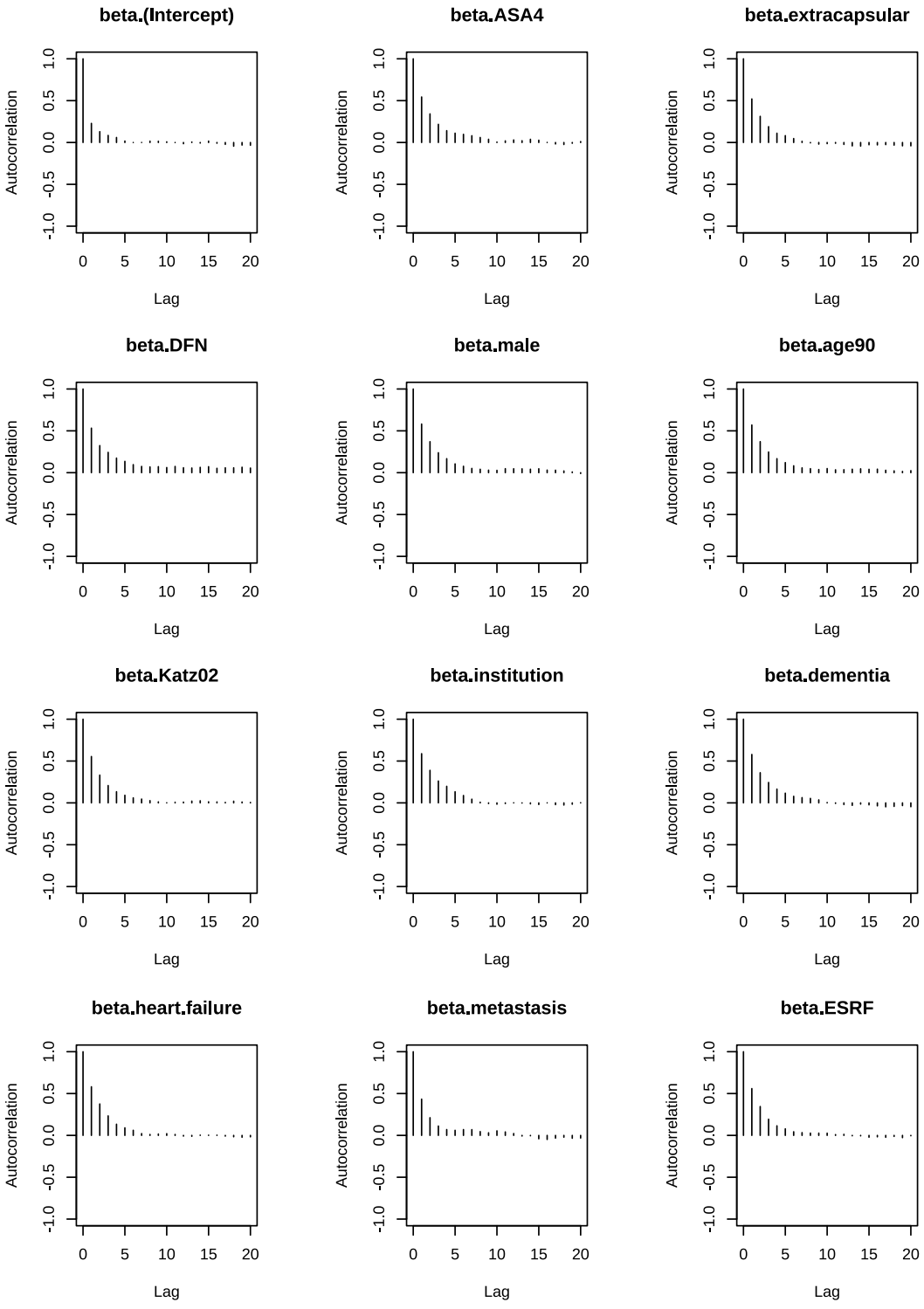

Figure B.5: Autocorrelation plots of Markov chains for each β-coefficient.

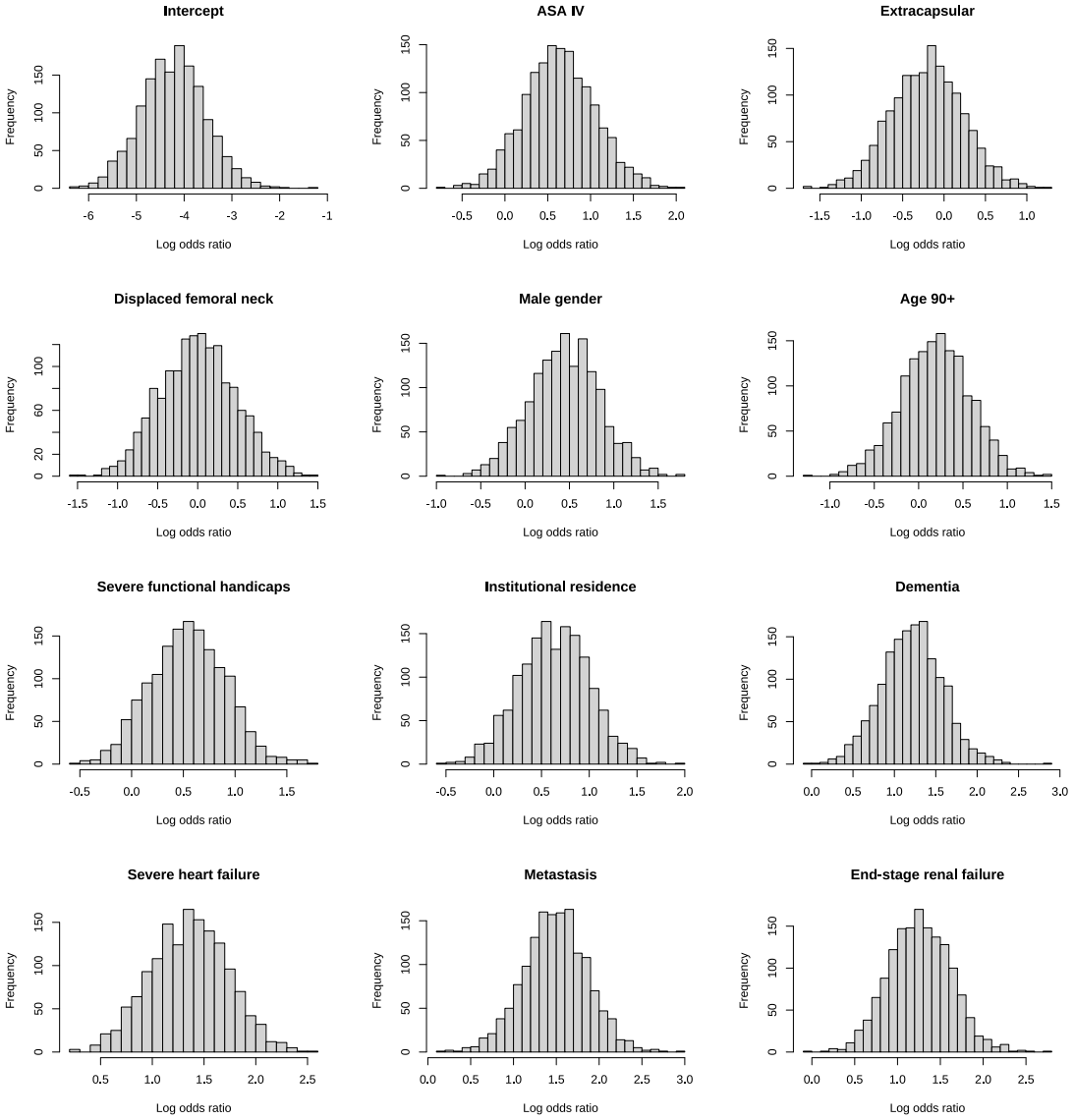

Figure B.6: Histograms depicting the distributions of posterior samples drawn through Markov Chain Monte Carlo, for each β-coefficient.

APPENDIX C

Activity Recognition Supplements

C.1 Rationale Behind Chosen Number of Sensors

The static activities of interest for the human activity recognition (HAR) system included sitting, standing, and lying down. Based on the analysis of expected accelerometer values for these activities, as shown in Table C.1, it was postulated that a single accelerometer was unable to distinguish between all static activities. The MOX on the upper leg was anticipated to yield highly similar axes orientations during sitting and lying down (supine). The APDM on the lower back was anticipated to yield highly similar axes orientations during sitting and standing. Combined, however, sufficient distinct information should be available to distinguish between all static activities. Therefore, a minimum of two accelerometers was expected to be necessary to successfully develop a robust HAR system.

Table C.1: Overview of expected accelerometer values (m/s^2) for the static activities of interest for the human activity recognition task. The expected accelerometer values are provided for the MOX on the upper leg, and the APDM on the lower back.

Static activities	**MOX orientation (upper leg)**			**APDM orientation (lower back)**		
	X-axis	Y-axis	Z-axis	X-axis	Y-axis	Z-axis
Sitting[ab]	0	0	9.81	-9.81	0	0
Standing[b]	9.81	0	0	-9.81	0	0
Lying down (supine)[a]	0	0	9.81	0	0	9.81
Lying down (facing left)	0	9.81	0	0	-9.81	0
Lying down (facing right)	0	-9.81	0	0	9.81	0

[a] It is anticipated that the MOX alone is unable to distinguish between distinguish between sitting and supine
[b] It is anticipated that the APDM alone is unable to distinguish between sitting and standing

C.2 Anomalous Drifts in Walking Accelerations

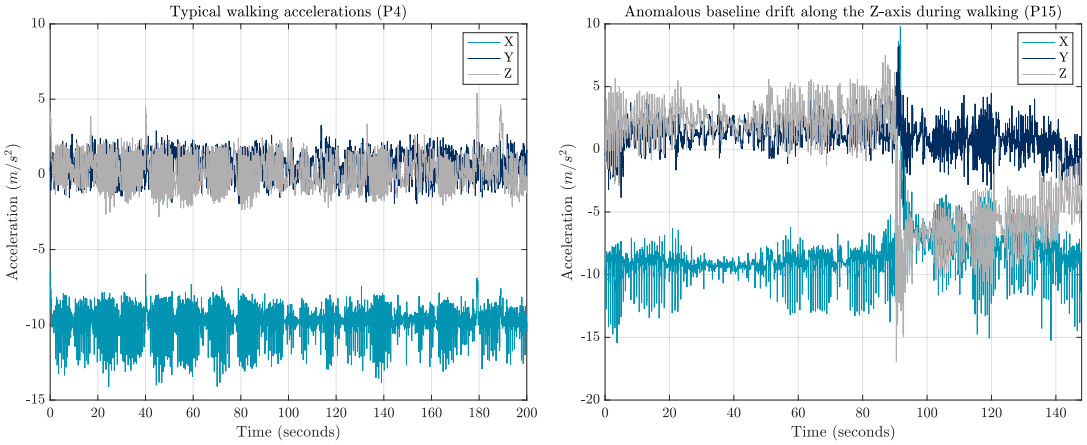

Figure C.1: Comparison of typical walking accelerations (left panel), and anomalous walking accelerations with a baseline drift along the Z-axis (right panel).

C.3 Individual Participant Performance

Table C.2: Individual participant performance for each activity. The F1-scores are reported for the feature intervention model (FIM), control condition model (CCM) and complete data intervention model (CDIM).

Participant	F1 Walking			F1 Standing			F1 Sitting			F1 Lying down			F1 Transfer		
	FIM	CCM	CDIM	FIM	CCM	CDIM	FIM	CCM	CDIM	FIM	CCM	CDIM	FIM	CCM	CDIM
1	0.45	0.11	1.00	0.86	0.88	0.93	1.00	1.00	0.96	1.00	1.00	1.00	0.50	0.18	0.15
2	0.83	0.85	0.69	0.93	0.94	0.94	1.00	0.99	0.93	1.00	0.98	1.00	0.67	0.58	0.18
3	0.89	0.87	0.82	0.91	0.89	0.92	1.00	1.00	0.67	0.99	1.00	1.00	0.76	0.76	0.04
4	0.95	0.93	0.93	0.97	0.94	0.94	1.00	0.99	0.99	1.00	1.00	1.00	0.90	0.90	0.90
5	0.93	0.87	1.00	0.96	0.93	0.95	1.00	1.00	0.33	1.00	1.00	1.00	0.97	0.84	0.04
6	0.90	0.96	0.60	0.90	0.95	0.91	1.00	1.00	0.91	0.94	1.00	0.94	0.70	0.77	0.10
7	0.92	0.97	0.86	0.94	0.98	0.96	0.99	0.99	0.98	1.00	1.00	1.00	0.72	0.67	0.57
8	0.90	0.90	0.67	0.88	0.85	0.87	1.00	0.99	0.96	0.99	0.99	0.99	0.69	0.64	0.31
9	0.91	0.71	0.78	0.96	0.90	0.94	0.99	1.00	1.00	1.00	1.00	1.00	0.71	0.75	0.88
10	0.90	0.91	1.00	0.93	0.93	0.94	1.00	1.00	0.99	1.00	1.00	1.00	0.88	0.78	0.69
11	0.90	0.92	0.80	0.93	0.95	0.93	1.00	1.00	1.00	0.97	0.97	0.97	0.89	0.86	0.86
12	0.93	0.88	0.62	0.97	0.95	0.97	0.99	0.99	0.99	1.00	1.00	0.99	0.77	0.61	0.62
13	0.89	0.93	0.86	0.88	0.92	0.88	0.98	0.99	1.00	0.93	0.99	0.00	0.83	0.67	0.12
14	0.92	0.92	0.78	0.97	0.97	0.97	1.00	1.00	0.37	0.91	0.91	0.91	0.80	0.88	0.03
15	0.90	0.88	0.86	0.95	0.95	0.94	1.00	1.00	0.51	1.00	1.00	0.99	1.00	0.92	0.04
16	0.94	0.80	0.92	0.94	0.85	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.86	0.92
17	0.95	0.89	0.73	0.97	0.95	0.96	1.00	1.00	1.00	0.91	1.00	0.91	0.85	0.70	0.85
18	0.87	0.84	0.79	0.91	0.86	0.93	0.99	0.99	1.00	0.00	0.15	0.00	0.72	0.77	0.59
19	0.92	0.96	0.67	0.92	0.93	0.93	1.00	1.00	0.94	0.97	0.97	0.99	0.80	0.72	0.27
20	0.95	0.88	0.87	0.83	0.68	0.86	1.00	1.00	1.00	0.94	1.00	1.00	0.85	0.80	0.93
21	0.97	1.00	0.93	0.95	1.00	0.95	1.00	1.00	0.93	1.00	1.00	1.00	0.86	0.89	0.20
22	0.92	0.92	0.89	0.93	0.95	0.96	1.00	1.00	0.86	1.00	1.00	1.00	1.00	0.91	0.37
23	0.93	0.84	0.93	0.97	0.91	0.97	1.00	1.00	0.90	1.00	1.00	1.00	0.97	0.93	0.16
24	0.92	0.91	0.85	0.88	0.82	0.80	1.00	0.99	1.00	0.95	0.95	0.91	0.86	0.74	0.83