# PREDICTING THE BLEEDING RISK IN PEOPLE WITH HEMOPHILIA

# PREDICTION OF THE RISK OF BLEEDING IN PEOPLE LIVING WITH HEMOPHILIA

By FEDERICO GERMINI, MD, MSc

A thesis submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the PhD Degree in Health Research Methodology.

McMaster University © Copyright by Federico Germini, April 2023

McMaster University PhD in Health Research Methodology (2022) Hamilton, Ontario

# TITLE: Prediction of the Risk of Bleeding in People Living with Hemophilia

## AUTHOR: Federico Germini, MD MSc

**SUPERVISOR:** Professor A. Iorio

NUMBER OF PAGES: xi, 253

## Lay abstract

People living with hemophilia lack a coagulation factor and tend to experience spontaneous bleeds, with frequency and intensity that vary between individuals. Predicting who will experience more bleeds would allow for changing the treatment strategies and directing the best resources to the persons that can benefit more.

Through this project, we identified the variables that should be considered to estimate the risk for bleeding in people living with hemophilia, namely the blood levels of the lacking coagulation factor, the bleeding history, the physical activity levels, the concomitant treatment with blood thinners, and the presence of obesity. We determined that Fitbit Charge and Charge HR are the most accurate devices for measuring steps and Apple Watch for heart rate. Lastly, we found that an existing tool for predicting the risk of bleeding is not accurate enough to be used in this setting, and a new model should be produced.

# Abstract

A tool allowing the prediction of the risk of bleeding in patients with hemophilia would be relevant for patients, stakeholders, and policymakers.

We performed a systematic review of the literature searching for available risk assessment models to predict the risk of bleeding in people living with hemophilia, and to determine the key risk factors that the ideal model should include. We also systematically review the literature to determine the acceptability and accuracy of wrist-wearable devices to measure physical activity in the general population. Finally, we validated the performance of a risk assessment model for the prediction of the risk for bleeding in people living with hemophilia.

We identified the following risk factors for bleeding in people living with hemophilia: plasma factor levels, history of bleeds, physical activity, antithrombotic treatment, and obesity. The FitBit Charge and FitBit Charge HR are the most accurate devices for measuring steps, and the Apple Watch is the most accurate for measuring heart rate. No device proved to be accurate in measuring energy expenditure. The predictive accuracy of the risk assessment model that we validated does not endorse its use to drive decision making on treatment strategies based on the predicted number of bleeds. This might in part be explained by the methods used in the derivation phase.

The need for an accurate risk assessment model to predict the risk of bleeding in people living with hemophilia is still unmet. This should be done by including the relevant risk factors identified through our work, with data on physical activity possibly collected using an accurate wrist-wearable device, and through the application of rigorous methods in the derivation and validation phases.

## Acknowledgements

Since I arrived at McMaster University, I have been working in a collaborative, stimulating environment. Three mentors have been crucial in this journey. One was running in front of me, dropping plenty of opportunities for me to pick up, and keeping me fit. One took me by the hand and accompanied me since when my first application to the HRM program was rejected. One pushed me toward places I would have not explored. All of this was essential to get where I am now. I hope to continue to benefit from such an excellent company and treasure this experience to give back to others. Being Italian, I also must thank my mom, dad, and my truly supportive, understanding, family.

# Table of Contents

*Lay abstract* ..... *iii*

*Abstract* ..... *iv*

*Acknowledgements* ..... *v*

*Table of Contents* ..... *vi*

*Lists of Figures and Tables* ..... *ix*

*List of all Abbreviations and Symbols* ..... *x*

*Declaration of Academic Achievement* ..... *xi*

*Chapter 1 – Introduction* ..... *1*

**Objectives** ..... **6**

**Figures** ..... **7**

**References** ..... **8**

*Chapter 2 – Risk factors for bleeding in hemophilia* ..... *11*

**Authors** ..... **11**

**Affiliations** ..... **11**

**Corresponding author** ..... **12**

**Essentials** ..... **13**

**Abstract** ..... **14**

**Keywords** ..... **15**

**Background** ..... **16**

**Objectives:** ..... **16**

**Methods** ..... **17**

**Results** ..... **21**

**Discussion** ..... **26**

**Conclusions** ..... **28**

**Funding** ..... **28**

**Conflict of Interest** ..... **28**

**Authors' contributions** ..... **29**

**References** ..... **30**

**Figures** ..... **34**

**Tables** ..... **36**

***Chapter 3 – Measuring physical activity*** ..... ***44***

**Abstract** ..... **44**

**Background** ..... **46**

**Methods** ..... **49**

**Results** ..... **53**

**Discussion** ..... **83**

**Conclusions** ..... **88**

**Declarations** ..... **89**

**References** ..... **90**

***Chapter 4 – validating a possible predictive model*** ..... ***103***

**Authors** ..... **103**

**Affiliations** ..... **103**

**Corresponding author** ..... **104**

**Background** ..... **105**

**Objectives:** ..... **106**

**Methods** ..... **107**

**Results** ..... **112**

**Discussion** ..... **114**

**Conclusions** ..... **116**

**Funding** ..... **116**

**Authors' contribution** ..... **116**

**References** ..... **117**

**Figures** ..... **118**

**Tables** ..... **122**

***Chapter 5 – Discussion, future directions, and conclusions.*** ..... ***124***

**References** ..... **129**

***Supplementary material - Chapter 2*** ..... ***131***

***Supplementary material – Chapter 3*** ..... ***134***

***Supplementary material – Chapter 4*** ..... ***153***

***Supplementary material – Chapter 5*** ..... ***238***

# Lists of Figures and Tables

Figure 1-1: example of a PK estimate obtained using WAPPS.
Figure 1-2: WAPPS clinical calculator.
Figure 2-1: PRISMA Flow Diagram.
Figure 3-1: PRISMA Flow Diagram.
Figure 3-2: summary of the results for the main accuracy outcomes.
Figure 4-1: observed versus predicted number of bleeds during each treatment period, for PWH A and an individual PK available.
Figure 4-2: observed versus predicted survival time (first bleed) during each treatment period, for PWH A and an individual PK available.
Figure 4-3: observed versus predicted number of bleeds during each treatment period, for PWH A without an individual PK available.
Figure 4-4: observed versus predicted survival time (first bleed) during each treatment period, for PWH A without an individual PK available.
Table 2-1: characteristics of the included studies.
Table 2-2: Risk of bias assessment of the included studies, using the QUIPS tool.
Table 2-3: risk factors, outcomes definitions, risk estimates, and covariates adjusted for.
Table *3*-1:4 Characteristics of the studies reporting on accuracy.
Table 3-2: 5 Characteristics of the studies reporting on acceptability.
Table 3-3: Result characteristics of the studies reporting on accuracy.
Table 4-6: Parameter estimates for the risk assessment model.
Table 4-7: Characteristics of the population and outcomes distribution.
Table 4-8: regression for observed versus predicted number of bleeds.

# List of all Abbreviations and Symbols

| |
|---|
| ABR: annualized bleeding rate |
| aIIR: adjusted incidence rate ratio |
| aOR: adjusted odds ratio |
| Apps: applications |
| ATHN: American Thrombosis and Hemostasis Network |
| AUROC: area under the receiver characteristics curve |
| BMI: body mass index |
| BSV: between subject variability |
| CBDR: Canadian Bleeding Disorders Registry |
| CDSR: Cochrane Database of Systematic Reviews |
| CENTRAL: Cochrane Central Register of Controlled Trials |
| CI: confidence interval |
| EE: energy expenditure |
| EHL: extended half-life |
| ETP: endogenous thrombin potential |
| FIX: factor IX |
| FVIII: factor VIII |
| HIREB: Hamilton Integrated Research Ethics Board |
| ICC: intraclass correlation coefficient |
| IU: international units |
| MAPE: mean absolute percentage |
| mHealth: mobile health |
| MVPA: mean to vigorous physical activity |
| OR: odds ratio |
| PAC-1: procaspase-activating compound 1 |
| PK: pharmacokinetics |
| PMCH: Personalized Medicine for Canadians with Hemophilia |
| PRISMA-ScR: PRISMA extension for Scoping Reviews |
| PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| PSSUQ: Post-Study System Usability Questionnaire |
| PWH: people living with hemophilia |
| QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies – version 2 |
| RAM: risk assessment model |
| ROC: receiver operating characteristic |
| RTTE: repeated time to event |
| SD: standard deviation |
| SHL: standard half-life |
| UKHCDO: United Kingdom Haemophilia Centre Doctors' Organisation |
| WAPPS-Hemo: Web-Accessible Population Pharmacokinetic Service – Hemophilia |

## Declaration of Academic Achievement

This project originated from a discussion initiated by my supervisor, Alfonso Iorio. We worked together to develop it, and I took the lead from there. People's contribution to specific parts of this project is reported in each chapter.

# Chapter 1 – Introduction

Hemophilia is an inherited X-linked bleeding disorder. Hemophilia A is characterized by a deficiency of clotting factor VIII, while factor IX is deficient in hemophilia B. The cornerstone of hemophilia care is the treatment with the deficient clotting factor, although treatment alternatives start being available.[1] The severity of bleeding episodes is usually associated with clotting factor levels. Hemophilia with factor levels of <0.01 IU/ml is classified as severe. Severe hemophilia is associated with spontaneous bleeds into joints or muscles, even in the absence of identifiable hemostatic challenges.[1] In untreated people with severe hemophilia, recurrent bleeds in joints progressively cause disabling arthritis. On the other hand, these spontaneous bleeds seldom occur in people living with hemophilia (PWH) with factor levels >0.01 IU/ml.[2] This observation led to the introduction of regular prophylaxis, i.e. the regular treatment with factor concentrates intending to maintain the factor levels above a certain threshold. Prophylactic treatment is usually dosed by weight (with higher or lower doses according to different protocols), but due to high inter-patient variability in the drug pharmacokinetics (PK), this can translate into either under- or over-dosing.[3] This can lead to a waste of resources or low efficacy. For this reason, the use of a tailored treatment approach based on the individual PK profile has been advocated, having the potential to be superior to weight-based regimens in terms of efficacy and resource consumption.[4,5] The Web-Accessible Population Pharmacokinetic Service - Hemophilia (WAPPS-Hemo, NCT02061072) is a population-based Bayesian calculator that provides caregivers with individual patients' PK estimates for many factor concentrates.[6] These population-based PK estimates are calculated on a minimum of two post-infusion blood

samples, as compared to the 9 to 11 samples needed for a classic individual PK.[3,6,7] Thanks to the Bayesian approach, the variability across different segments of the PK curve can vary at different time points, depending on the availability of information from the individual (smaller variability) or only from the population (larger variability).[8] The service is hosted at McMaster University, is industry independent, and is freely available online.[9] An example of a PK estimate is provided in **Figure 1-2**. Once the individual PK is assessed, the service provides the user with a calculator that, given two out of the three following variables: dose, administration frequency, or hematic factor trough levels, calculates the third one (see **Figure 1-3**). This allows the physician to perform simulations on the effect of different treatment plans, in terms of dosage and frequency of administration, or to calculate the dose and frequency needed to keep the factors trough levels above a specified value. The PK based approach for tailoring prophylaxis regimens is gaining popularity and has been also suggested (possibly using WAPPS-Hemo) by the United Kingdom Haemophilia Centre Doctors' Organisation (UKHCDO).[10] The effects of the application of a PK-based approach on resource consumption and clinical outcomes are being studied in the Personalized Medicine for Canadians with Hemophilia (PMCH) study, a pragmatic multi-center study with a before-after design, on the clinical impact of the use of the WAPPS-Hemo service to tailor the factor replacement prophylaxis strategy in Canadian people with hemophilia (PWH) A or B (NCT03615053). The PMCH study is producing a large amount of data, collected in the Canadian Bleeding Disorders Registry (CBDR), a national clinical database that collates clinical information, clotting factor utilization, and patient outcomes on people with bleeding disorders. Factor utilization is reported by PWH daily, in real time. One possible use of these data is the prediction of the patients' risk of bleeding. PWH,

physicians, and policymakers might benefit from knowing the risk of bleeding in individual patients. From the patients' perspective, it would be helpful to know what's their risk of bleeding and how this might change modifying potential risk factors. For example, knowing that changing the treatment adherence from 70 to 90% would reduce the annualized risk of bleeding of a certain amount, might help a patient in improving his adherence. Moreover, knowing the punctual risk of bleeding based on risk factors including modifiable variables such as the plasma level concentration and the type of physical activity to be performed, would allow the patients to change their risk, modifying the factor plasma levels (with an extra infusion) before a high-risk activity, or avoiding high-risk activities when the factor levels are too low. On the other hand, when the risk for bleeding is very low, a person with hemophilia might reduce the factor usage, and this would allow for saving resources. From the physicians' perspective, a risk assessment model (RAM) could be used for educational purposes as described above, and to select the best treatment for a specific PWH, for example reserving more resource-intense treatment regimens for PWH at high risk of bleeding. From a policy-maker perspective, the identification of different risk categories would allow them to decide how to allocate resources. This is particularly important now that new therapies entered or are about to enter the market. Emicizumab is a humanized antibody that mimics the function of factor VIII and presents potential clinical advantages as compared to factor concentrates, being administered subcutaneously instead of intravenously, and less frequently.[11,12] Another option will soon be available: gene therapy for PWH A and B have been tested in phase 1 and 2 trials with promising results[13,14], several other phase 2 studies are ongoing and some company already moved to phase 3 (NCT03392974, NCT03370913). It is reasonable to assume that these therapies

will be very expensive, and that policymakers will have to decide on the amount of resources to allocate to these treatments, and which groups of patients will be eligible to receive them. The identification of patients at high risk of bleeding might be a way to select the patients that can benefit more from the new treatments.

We are not aware of the existence of any RAM for the estimation of the risk of bleeding in PWH. However, to avoid reinventing the wheel, one should perform a systematic review of the literature to confirm this. If this is confirmed, the following natural step would be the derivation of a RAM, its internal and external validation, and its use in a management study, to show that clinically relevant outcomes can be positively impacted using the RAM.[15] Ideally, such a RAM should include all the relevant known risk factors. The best way to identify such risk factors and to estimate the strength of the association between each risk factor and the outcome is again a systematic review of the literature.[16] Two of these risk factors are the patients' plasma factor levels and exposure to physical activity.[17] Regarding the former, the use of infusion data from CBDR and PK data from WAPPS would allow us to estimate the time-varying plasma factor levels of PWH. As per physical activity, at the present moment, no information is systematically collected in CBDR or WAPPS. This might be done with a questionnaire aimed at classifying the level of physical activity.[10] Problems with this approach would be that the measure is subjective, the temporal relationship between the level of physical activity and the bleeding episodes would be hard to determine, and data might be missing when a patient is not willing to complete the questionnaire. Another option would be to passively collect data on physical activity using a wearable device (e.g., a smartwatch). We think that this option would be preferable, for the following reasons: 1) the patients would be required the only effort of wearing the device,

2) data would be potentially more precise and richer, and 3) the temporal relation with treatments and bleeds would be easier to establish. We did not find any published or ongoing study aimed at measuring physical activity using wearable devices in persons living with hemophilia. The first step toward measuring physical activity using a wearable device would be the identification of a device that is accurate and that users found acceptable. Many studies have been conducted to assess the accuracy and acceptability of one or more wearable device, and a few systematic reviews on the topic has been published.[18,19] However, we found no comprehensive systematic review on the accuracy and acceptability of wrist-wearable devices, without limitations in terms of device models and type of physical activity measured.

Going back to factor levels as a predictor of the bleeding risk, our group, led by a PK modeler, derived a model to assess the relationship between factor VIII activity levels and bleeding risk, using data from CBDR and the American Thrombosis and Hemostasis Network (ATHN). The paper describing the model derivation is reported in the supplementary material. This model was not derived with the specific intent of predicting the risk for bleeding, but could be used for this purpose, and seemed to function well in the derivation cohort. However, as mentioned above, the performance of such a model would need to be validated internally and externally.[15] If the good predictive performance of this model were confirmed, there would be no need to derive a new multivariable RAM. The internal validation can be performed using CBDR data that has been collected after the model derivation.

## Objectives

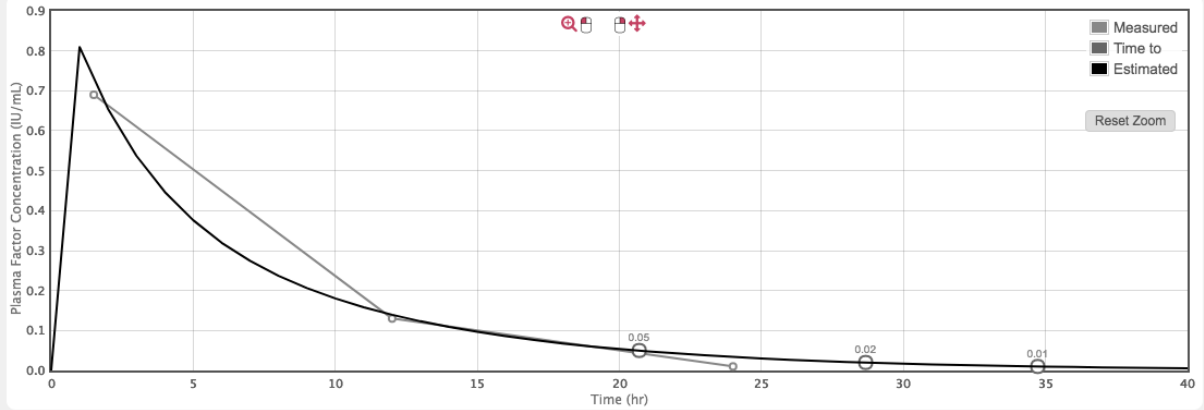
Objectives of this project were the following:

1. To conduct a systematic review of RAMs for predicting the risk of bleeding in PWH;
2. In the absence of existing RAMs, to conduct a systematic review of risk factors for bleeding in PWH;
3. To conduct a systematic review on the accuracy and acceptability of wrist-wearable activity tracking devices for measuring physical activity;
4. To validate the performance of a recently derived RAM for the prediction of the risk of bleeding in PWH.

Objectives one and two above are the subject of the first paper reported in this thesis (Chapter 2). Objectives three and four are the subject of the second (Chapter 3) and third (Chapter 4) paper, respectively. The first two papers have been already published in per reviewed journals,[20,21] the third one will be submitted for publication soon after the submission of the derivation study (which will happen in Q1 2023, to align the publication of the study with its presentation to a conference). Chapter 5 reports a protocol for the derivation and internal validation of a new RAM that includes all the relevant risk factors identified through our systematic review of the literature, and the conclusions of this thesis project.

# Figures

*Figure 1-2: example of a PK estimate obtained using WAPPS*




**Infusion data used for this estimate:**

| +/- | ID | Drug | Height (cm) | BW (kg) | Tot IU | IU/kg | End of infusion | Infusion Duration (min) | Opt. | Notes |
|---|---|---|---|---|---|---|---|---|---|---|
| | 229 | Kogenate | N/A | 21 | 1250 | 59.5 | 2015-03-02 08:00 | 0 | | |

**Pharmacokinetic Estimate**

Please note that the **distance** between the *conservative* and the *optimistic* estimate is **directly proportional** to the uncertainty of the estimation. When the values are distant enough that the choice between a *balanced* or *conservative* estimate would result in a different clinical decision (e.g. they are more than 24 hrs apart, suggesting an infusion regime one day shorter), we recommend using the *conservative* estimate.

| | Estimate | | |
|---|---|---|---|
| **Parameter** | **Conservative** | **Balanced** | **Optimistic** |
| Time to 0.05 IU/ml (hr) | 17.25 | 20.75 | 24.00 |
| Time to 0.02 IU/ml (hr) | 23.75 | 28.75 | 33.75 |
| Time to 0.01 IU/ml (hr) | 28.50 | 34.75 | 41.00 |
| Half-life (hr) | 4.75 | 6.00 | 7.25 |

**Concentrate Info**

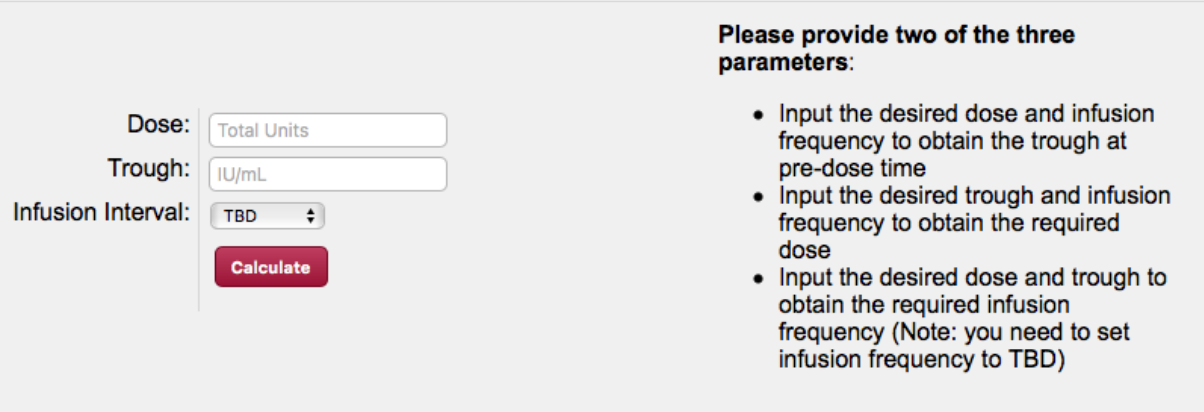
| Concentrate | Lab Test | Standard |
|---|---|---|
| Kogenate | One Stage Coag. (PTT Based) | Drug Specific |

**Patient Data**

| | |
|---|---|
| **Age** | 9 |
| **Body Weight (kg)** | 21.0 |
| **Baseline Factor Level (IU/mL)** | < 0.010 |
| **Gender** | M |
| **Blood Group** | O |

BW: body weight; IU: international units; M: male; N/A: not available, PTT: partial thromboplastin time.

*Figure 1-3: WAPPS clinical calculator*




IU: international units; TDB: to be determined.

## References

1 Srivastava A, Brewer AK, Mauser-Bunschoten EP, *et al.* Guidelines for the management of hemophilia. *Haemophilia* 2013;**19**:e1-47. doi:10.1111/j.1365-2516.2012.02909.x

2 Arnold WD, Hilgartner MW. Hemophilic arthropathy. Current concepts of pathogenesis and management. *J Bone Joint Surg Am* 1977;**59**:287–305.

3 Hazendonk HCAM, Lock J, Mathôt RAA, *et al.* Perioperative treatment of hemophilia A patients: blood group O patients are at risk of bleeding complications. *Journal of Thrombosis and Haemostasis* 2016;**14**:468–78. doi:10.1111/jth.13242

4 Björkman S, Carlsson M. The pharmacokinetics of factor VIII and factor IX: Methodology, pitfalls and applications. Haemophilia. 1997;**3**:1–8. doi:10.1046/j.1365-2516.1997.00074.x

5 Collins PW, Fischer K, Morfini M, *et al.* Implications of coagulation factor VIII and IX pharmacokinetics in the prophylactic treatment of haemophilia. Haemophilia. 2011;**17**:2–10. doi:10.1111/j.1365-2516.2010.02370.x

6 Iorio A, Keepanasseril A, Foster G, *et al.* Development of a Web-Accessible Population Pharmacokinetic Service-Hemophilia (WAPPS-Hemo): Study Protocol. *JMIR Res Protoc* 2016;**5**:e239. doi:10.2196/resprot.6558

7 Brekkan A, Berntorp E, Jensen K, *et al.* Population pharmacokinetics of plasma-derived factor IX: procedures for dose individualization. *Journal of Thrombosis and Haemostasis* 2016;**14**:724–32. doi:10.1111/jth.13271

8 McEneny-King A, Foster G, Iorio A, *et al.* Data Analysis Protocol for the Development and Evaluation of Population Pharmacokinetic Models for Incorporation

Into the Web-Accessible Population Pharmacokinetic Service - Hemophilia (WAPPS-Hemo). *JMIR Res Protoc* 2016;**5**:e232. doi:10.2196/RESPROT.6559

9 WAPPS-Hemo (Web Accessible Population Pharmacokinetic Service – Hemophilia ). https://www.wapps-hemo.org/ (accessed 7 Oct 2021).

10 Collins P, Chalmers E, Chowdary P, *et al.* The use of enhanced half-life coagulation factor concentrates in routine clinical practice: guidance from UKHCDO. *Haemophilia* 2016;**22**:487–98. doi:10.1111/hae.13013

11 Shima M, Hanabusa H, Taki M, *et al.* Factor VIII–Mimetic Function of Humanized Bispecific Antibody in Hemophilia A. *New England Journal of Medicine* 2016;**374**:2044–53. doi:10.1056/NEJMoa1511769

12 Mahlangu J, Oldenburg J, Paz-Priel I, *et al.* Emicizumab Prophylaxis in Patients Who Have Hemophilia A without Inhibitors. *N Engl J Med* 2018;**379**:811–22. doi:10.1056/NEJMoa1803550

13 Rangarajan S, Walsh L, Lester W, *et al.* AAV5–Factor VIII Gene Transfer in Severe Hemophilia A. *New England Journal of Medicine* 2017;**377**:2519–30. doi:10.1056/NEJMoa1708483

14 George LA, Sullivan SK, Giermasz A, *et al.* Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. *N Engl J Med* 2017;**377**:2215–27. doi:10.1056/NEJMoa1708538

15 Steyerberg EW, Moons KGM, van der Windt DA, *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;**10**. doi:10.1371/JOURNAL.PMED.1001381

16 Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *Br J Cancer* 2009;**100**:1219–29. doi:10.1038/SJ.BJC.6604999

17 Broderick CR, Herbert RD, Latimer J, *et al.* Association between physical activity and risk of bleeding in children with hemophilia. *JAMA - Journal of the American Medical Association* 2012;**308**:1452–9. doi:10.1001/jama.2012.12727

18 Feehan LM, Geldman J, Sayre EC, *et al.* Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth. 2018;**6**. doi:10.2196/10527

19 Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. International Journal of Behavioral Nutrition and Physical Activity. 2015;**12**. doi:10.1186/s12966-015-0314-1

20 Germini F, Noronha N, Debono VB, *et al.* Accuracy and Acceptability of Wrist-Wearable Activity-Tracking Devices: Systematic Review of the Literature. *J Med Internet Res* 2022;**24**. doi:10.2196/30791

21 Germini F, Noronha N, Abraham Philip B, *et al.* Risk factors for bleeding in people living with hemophilia A and B treated with regular prophylaxis: A systematic review of the literature. *J Thromb Haemost* 2022;**20**:1364–75. doi:10.1111/jth.15723

# Chapter 2 – Risk factors for bleeding in hemophilia

# Title:

# Risk factors for bleeding in people living with Hemophilia A and B treated with regular prophylaxis: a systematic review of the literature.

## Running head:

## Review of risk factors for bleeding in hemophilia.

## Authors

Federico Germini,[1,2] Noella Noronha[1], Binu Abraham Philip[1], Omotola Olasupo[1], Drashti Pete[1], Tamara Navarro[1], Arun Keepanasseril,[1] Davide Matino,[2] Kerstin de Wit,[1,2,4] Sameer Parpia,[1, 3] and Alfonso Iorio[1, 2].

## Affiliations

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Oncology, McMaster University, Hamilton, ON, Canada

[4]Department of Emergency Medicine, Queen's University, Kingston, ON, Canada

## Corresponding author

Federico Germini, MD MSc.

Department of Health Research Methods, Evidence and Impact.

Health Information Research Unit (HIRU), Communication Research Laboratory (CRL),

McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada:

germinif@mcmaster.ca

Main text word count: 3301

Number of references: 32

Number of tables and figures: 4

Abstract word count: 249

## Essentials

- Estimating the risk for bleeding in people with hemophilia (PWH) would be clinically relevant.
- We performed a systematic review on risk assessment models and risk factors for bleeding in PWH.
- No risk assessment model was found, but ten studies assessed possible risk factors.
- We identified some risk factors that should be considered when building such a model.

## Abstract

**Background:** Knowledge about the risk for bleeding in patients with hemophilia (PWH) would be relevant for patients, stakeholders, and policy makers.

**Objectives:** to perform a systematic review of the literature on risk assessment models (RAMs) and risk factors for bleeding in PWH on regular prophylaxis.

**Methods:** We searched MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials, and the Cochrane Database of Systematic Reviews from inception through August 2019. In duplicate, reviewers screened the articles for inclusion, extracted data, and assessed the risk for bias using the QUIPS tool. A qualitative synthesis of the results was not performed due to high heterogeneity in risk factors, outcomes definition and measurement, and statistical analysis of the results.

**Results:** From 1843 search results, 10 studies met the inclusion criteria. No RAM for the risk for bleeding in PWH was found. Most studies included only PWH A or both PWH A and B and were conducted in North America or Europe. Only one study had a low risk for bias in all the domains. Eight categories of risk factors were identified. The risk for bleeding was increased when factor levels were lower and in people with a significant history of bleeding or who engaged in physical activities involving contact.

**Conclusions:** Our findings suggest that plasma factor levels, history of bleeds, and physical activity should be considered for the derivation analysis when building a RAM for bleeding in PWH, and the role of other risk factors, including antithrombotic treatment and obesity, should be explored.

## Keywords

Bleeding

Hemophilia A

Hemophilia B

Hemorrhage

Risk Assessment

Risk Factors

## Background

The cornerstone of hemophilia care is the treatment with the deficient clotting factor.[1] Regular prophylaxis has been shown to be superior to episodic on-demand treatment for the prevention of bleeds and joint disease.[2] However, people with hemophilia (PWH) treated with similar regimens of regular prophylaxis can have very different bleeding patterns.[3] This is at least in part explained by the fact that prophylactic treatment is usually dosed by weight but, due to high inter-patient variability in the drug pharmacokinetics (PK), this can translate in either under- or over-dosing.[4] Other factors can have a role in determining the individual risk for bleeding, like physical activity levels,[5] bleeding history,[6] or treatment adherence.[7] These and other risk factors for bleeding could be combined in a risk assessment model (RAM) for the prediction of bleeds in PWH. Patients, physicians, and policymakers might benefit from knowing the risk of bleeding of individual patients. This is particularly important now that new non-clotting factor therapies (e.g., emicizumab) are available [8] or are about to enter the market (e.g., gene therapy).[9,10] The identification of patients at high risk of bleeding is an important first step in lowering the risk of bleeding. Also, modifying some of the identified risk factors serves to select patients that can benefit more from new treatments.

**Objectives:**

To perform a systematic review of the literature on RAMs and on risk factors for the prediction of the risk of bleeding in PWH treated with regular prophylaxis using clotting factor therapies.

## Methods

We conducted a systematic review of RAMs and risk factors for bleeding in PWH. The systematic review was designed referring to the CHARMS checklist for systematic reviews of prediction models[11] and the guide to systematic review and meta-analysis of prognostic factor studies from Riley et al.[12] The review was registered in PROSPERO (CRD42020152511).

**Criteria for selecting studies for this review**

**Types of studies:** The following studies were eligible for the systematic review: prognosis studies based on data from randomized controlled trials, cohort studies (both retrospective and prospective), registry-based studies, and case-control studies.

**Patients:** Studies on patients with congenital hemophilia A or B on regular prophylaxis were included. No age limit was applied. Studies not reflecting the general population of interest were excluded, such as studies limited to PWH treated on-demand, to patients with inhibitors, or focusing on hemophilia carriers.

**Exposure:** We investigated all the prognostic factors reported in the individual studies.

**Comparison:** The comparator was the absence or different levels of any risk factor.

**Outcome:** The primary outcome of the review was any bleeding episode, defined as an event interpreted as a bleed by the patients or his/her physician, and treated with a clotting factor concentrate.[13] We excluded studies focusing only on specific bleeding sites, e.g. intracranial bleeding or gastrointestinal bleeding, as risk factors for these events, such as modality of delivery and Helicobacter pylori infections, are unlikely to be generalizable to all bleeds.[14,15]

**Setting:** Outpatients, during everyday life. Studies in perinatal and perioperative settings were excluded.

**Time:** If available, we extracted data on the risk at 12 months, but this could vary from study to study.

**Search methods for identification of studies:**

We searched MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials (CENTRAL), and the Cochrane Database of Systematic Reviews (CDSR) from inception to August 21, 2019. The search strategy was built with the help of a librarian and is available in the supplementary material. No date and language limits were applied. We combined terms related to hemophilia, bleeding, and risk factors, including the highly sensitive search filter for identifying prognosis studies in MEDLINE and EMBASE.[16]

**Data collection and analysis**

**Selection of studies:** After a calibration exercise, the reviewers screened the titles and abstracts of the retrieved articles for inclusion, using prespecified inclusion and exclusion criteria, using the online software Rayyan.[17] They then screened the full texts of potentially eligible studies. Both phases of the screening process were conducted independently and in duplicate by two reviewers. Reviewers discussed disagreements, and a third senior reviewer was consulted to resolve them when necessary.

**Data extraction and management:** Two reviewers extracted data on the studies' characteristics and outcomes using a Microsoft Excel Spreadsheet. The data extraction form was prepared according to the CHARMS checklist[11] and based on inputs from experts in the field. Even in this case, we conducted a calibration exercise before starting the extraction. The extraction was performed independently in duplicate, with a third senior

reviewer resolving disagreements when necessary. For each included study, we collected data on the following characteristics: study context (country and year of publication), study design, population and demographics (inclusion and exclusion criteria, age, disease type, severity, and treatment, sample size), outcome (definition, method of measurement, duration of follow-up), risk factors (definition, method of measurement), statistical analysis (model used, methods for handling the outcome, the risk factors, adjustment for other predictors, and missing data), number of events in exposed and non-exposed cohorts, measures of association [e.g., hazard ratio and 95% confidence intervals (CIs)].

**Assessment of risk of bias in included studies:** We assessed the risk of bias of the included studies using the QUIPS tool, assessing the risk of bias in the domains of study selection, study attrition, measurement of the prognostic factors and outcomes, consideration of other predictors, and statistical analysis.[18] Regarding the consideration of other predictors, we pre-specified that studies would be assessed to be at high risk for bias if strategies to consider the effect of the following possible predictors were not implemented: severity of hemophilia, treatment type (prophylaxis vs on-demand), age, physical activity, and factor levels.

**Dealing with missing data:** We looked for information in clinical trials registries, study protocols, and secondary publications. Furthermore, we contacted the study authors to ask for missing data.

**Data synthesis:** Due to the significant heterogeneity observed between studies, in particular regarding the type of risk factors and outcomes included, how they were measured, and the risk estimates provided, we decided not to perform a quantitative synthesis, and provided a narrative synthesis of the results for both of our study objectives.

**Ethics**

Since this review used publicly available secondary data, no ethical approval was required.

## Results

Our search identified 1858 references. Among 152 full texts assessed, 11 articles from 10 studies met the inclusion criteria. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram[19] is reported in **Figure 2-4**. None of the retrieved studies aimed at developing or validating a risk assessment model, therefore we are only reporting studies on risk factors.

### Description of the included studies

**Table 2-9** describes the characteristics of the included studies. Five studies included only PWH A, one only PWH B, and four included both A and B. Severe hemophilia was included in all the studies, moderate in five, and mild in four. The study sample ranged from 32 to 286 individuals. The follow-up duration was most frequently 12 months, ranging from 6 to 72. Most of the studies were conducted in North America and/or Europe.

### Risk of bias assessment

**Table 2-10** reports the risk of bias assessment for each study. Only one study (Broderick et al[5]) was deemed to be at low risk of bias for all the domains. All the remaining studies were at high risk of bias in at least two domains.

### Prognostic factors for bleeding

The risk factors, outcome definitions, risk estimates, and co-variates considered in the analysis for each included study are reported in **Table 2-11**.

***Estimated plasma factor levels:*** 6 studies explored the association between plasma factor levels and the risk for bleeding. In these studies, the plasma factor levels were not directly measured, but based on individual or population-based PK estimates and the patients' treatment regimens or treatment diaries. Broderick et al, showed that, after adjusting for

physical activity levels and independently of individual characteristics, the adjusted odds ratio (aOR) for plasma factor levels was 0.98 (95% CI 0.97-0.99).[5] This can be interpreted as a 2% reduction in the risk for bleeding for every 1 IU/mL increase in the plasma factor levels. Collins et al, using individual PK estimates, calculated that, after adjusting for bleed cause, bleed site, age, and weight ratio (actual weight/ideal weight), for each additional hour per week spent with plasma FVIII levels <1 IU/mL the annual bleed rate increased by 2.2% (95% CI 1.6; 2.8) in PWH aged 1-6 years and 1.4% (95% 0.2; 2.6) in PWH aged 10-65 years.[7] In Ahnstrom et al, the time spent with plasma factor levels below 3, 2, and 1 IU/mL was not associated with overall bleeds and joint bleeds in a univariate analysis.[20] Fosser et al calculated that maintaining the FIX trough level >2, >5 and >10 IU/mL was associated with a 69% (95% CI 53; 80%), 77% (67; 84%), and 78% (69; 85%) reduction in the daily risk for bleeding, respectively.[21] Abrantes et al confirmed that the plasma FVIII concentration is inversely correlated with the bleeding risk, expressed as an annualized bleeding rate (ABR).[6] They used a Gompertz-survival model to estimate the effect of time-changing FVIII levels on the ABR. According to this model, keeping the FVIII levels constant at 5 IU/dL would reduce the ABR by ~1/3. Dargaud et al reported that the trough in the endogenous thrombin potential (ETP) was a better predictor of the occurrence of any spontaneous bleeding than the plasma FVIII concentration, with an area under the receiver characteristics curve (AUROC) of 0.94 (95%CI 0.88; 1.00) and 0.58 (95%CI 0.33; 0.84) respectively.[22]

***Treatment adherence/frequency:*** Collins et al estimated that, after adjusting for the bleeding cause and site, age, body habitus (measured as weight compared to ideal weight), and frequency of the prophylactic regimen, a 100% adherence to the prescribed treatment

regimen would have translated in 0.97 (95% CI 0.63; 1.27) fewer bleeds per year in the 1–6 years old and 1.19 (95% CI 0.66–1.61) fewer bleeds per year in the 10–65 years old.[7] In the study from Ross et al, the number of factor infusions per week was associated with an OR of 1.07 (95% CI 0.42; 2.73, per increase of one infusion) for having at least one bleed in an eight-week period.[23]

***Physical activity:*** Broderick et al[5] estimated the risk for bleeding associated with physical activity, categorized according to the American National Hemophilia Foundation.[24] This was a case cross-over study, therefore accounted for individual patients' characteristics, and the analysis was adjusted for the estimated plasma factor levels at the time of bleeding. The reference condition was inactivity and physical activity in which a collision is not expected (e.g., swimming). The aOR was 2.7 (95%CI 1.7; 4.8) for activities in which significant collisions might occur (e.g., basketball and gymnastics), and 3.7 (2.3; 7.3) for activities in which significant collisions are inevitable (e.g., rugby and martial arts). Ross et al estimated that practicing high-impact physical activities (corresponding to activities in which significant collision might occur or are inevitable, as per the American National Hemophilia Foundation) was associated with an aOR of 0.32 (0.04; 2.70) for having one or more bleed in an eight-week period.[23] The reference category was participating in activities in which a collision is not expected, and the risk estimate was adjusted for the frequency of prophylactic treatments.

***Bleeding history:*** In the study from Desjonqueres et al, a history of a non-severe bleed was a risk factor for severe bleed (i.e., a bleed requiring substitutive treatment, hospitalization, transfusion, or surgical/radiological intervention), with an OR of 21 (95% CI not reported, p 0.001)[25] Abrantes et al estimated that the number of bleeds in the previous 12 months

was a risk factor for bleeding: “Compared to a mean patient with 8.2 bleeds in the pre-study period, a patient who had one bleed (5th percentile of the observed data) or 84 bleeds (95th percentile) pre-study was found to have a 54% lower (95%CI 40-65) or 147% higher (95%CI 79-226) hazard, respectively.”[6]

***History of sport injuries:*** In the study from Ross et al, the number of injuries per trimester was associated with an OR of 7.02 (95% CI 0.30; 167, per increase of one injury) for having at least one bleed in an eight-week period. The definition of injury was not provided.[23]

***Age at target joint development:*** Gupta et al found an increased risk for joint bleeds in patients that developed a target joint after 5 years of age, as compared with developing a target joint before 5 years, with an adjusted incidence rate ratio (aIRR) ranging from 2.72 to 2.93 (95% CIs ranging between 2.04 and 2.93) across different age categories.[26]

***Age:*** In the study from Ross et al, the OR for having at least one bleed in an eight-week period was 1.04 (95% CI 0.81; 1.32) per every year increase of age.[23]

***Obesity:*** Gupta et al also explored the risk for joint bleeds associated with body mass index (BMI), reporting that, using normal/underweight PWH as a reference, the aIRR was 1.05 (95% CI 0.98; 1.13) in case of overweight, and 1.11 (1.04; 1.20) in case of obesity.[26]

***Antithrombotic treatment:*** Desjonqueres et al estimated that being on treatment with an antiplatelet agent, an anticoagulant, or both, was associated with a higher risk for bleeds requiring substitutive treatment, hospitalization, transfusion, or surgical/radiological intervention, with an OR of 3.55 (95% CI 1.2; 10.4).[25] No association was found between antithrombotic treatment and bleeds in general, independently of severity (no risk estimates or frequency distributions were reported for this outcome).

**Blood tests:** Jobe et al reported a correlation between the procoagulant platelet potential (ratio of procoagulant platelets and all activated platelets following stimulation with thrombin and convulxin) and the annualized bleeding rate in PWH (correlation coefficient $r = -0.47$, $p < 0.0001$).[27] They found no association with the P-selectin expression (granule release) or the procaspase-activating compound 1 (PAC-1) (risk estimates not reported).

**Seasonal variability:** Analyzing data from the same studies used by Collins, Fischer et al reported no seasonal variation in joint bleed rates in PWH A aged 1–6 years, while the occurrence of joint bleeds was increased in the summer, which accounted for 43% all joint bleeds for patients aged 10–17, and 46% for the ones 18–65 years.[28]

# Discussion

## Summary of findings

In our systematic review, we did not find any RAM for predicting the risk of bleeding in PWH. Only one of the studies reporting on risk factors for bleeding was at low risk of bias in all the domains, and the remaining studies were at high risk of bias in at least two domains. The between studies heterogeneity was high. All the studies but one agreed that higher plasma factor levels are associated with lower risk for bleeding, with variability in the risk estimates. This was confirmed across different techniques of estimating the plasma factor levels and using indirect measures like the treatment frequency and the adherence. Regarding the physical activity levels, the only study at low risk of bias on all the domains reported an increased risk of bleeding associated with physical activities involving collisions (95% CI of the aOR ranging from 1.7 to 7.3), while results were non-conclusive in another study. The bleeding history was also reported to be associated with the risk for bleeding, and in this case the different risk estimates pointed at a strong association. Other risk factors were only assessed in one study each, always at high risk for bias in at least two domains. These included age, BMI, antithrombotic therapy, season of the year, and some laboratory tests.

## Strengths and limitations

Benefits of our study include the rigorous methodology, the breadth of our search, and our duplicate and independent screening, data abstraction, and risk of bias assessment process. Another strength of our study is the coverage of studies on both RAMs and individual risk factors. The fact that we did not find any RAM is valuable information. Due to the presence of important heterogeneity between studies, we were not able to perform a meta-analysis.

Moreover, some studies had to be excluded because of missing information. To address these limitations, we contacted the authors and, in some cases, we were able to obtain additional information, but this did not allow us to fully address the issue. Since our review focused on PWH on regular prophylaxis, we excluded studies only including patients treated on-demand or patients with inhibitors. Therefore, our results should not be generalized to these populations. The risk of bias of the included studies was overall high. This is not a limitation of the review per se, but it affects the reliability of the results and needs to be considered.

**Future research directions**

More large high-quality studies with clearly defined predictors and outcomes (bleeding) in PWH are needed to identify additional clinical risk factors and biomarkers, and to validate the results of previous studies. A RAM for the estimation of the risk for bleeding is not available in the literature. For the reasons outlined in the introduction, we believe that deriving and validating such a RAM would be of value. Factor levels, physical activity, and bleeding history appear to be the most important risk factors to be included in such a model. However, these data are not conclusive, and the role of other risk factors should also be explored. This should be done on a population sampled appropriately and adjusting for important risk factors, as available studies were often flawed using convenience sampling and the lack of adjustment for important covariates. Factor levels should be estimated based on individual PK profiles and treatment logs. The use of treatment plans is a proxy for the logs and might lead to imprecise estimates. Measuring physical activity can be a burden for patients and clinicians/researchers, and information collected through questionnaires might be unreliable.[29,30] The use of activity trackers (e.g. smart watch) to collect this

information might help in addressing this issue. This would generate a large amount of data and their use for predicting the risk for bleeding would probably benefit from machine learning techniques. Lastly, efforts are needed to reduce the heterogeneity in the measurement and reporting in the field of risk for bleeding in hemophilia. Important sources of heterogeneity were the definitions of and methods for measuring the risk factors, the way these were handled in the statistical analysis, and the model used for the statistical analysis, yielding to heterogenous risk measures. A harmonized approach to measurement, handling, and reporting of these data is needed.

## Conclusions

Based on our systematic review of the literature, no RAM for the prediction of the risk for bleeding in PWH is available. Plasma factor levels, physical activity, and bleeding history are important risk factor for bleeding and ideally should be considered in the derivation of a RAM. The role of other risk factors, including antithrombotic treatment and obesity, should also be explored in the derivation process.

## Funding

Federico Germini received funding for this study from the Physicians' Services Incorporated (PSI) foundation through the PSI Research Trainee Fellowship.

## Conflict of Interest

Federico Germini's institution received research funding from NovoNordisk, Roche, Takeda, Bayer, Pfizer, and BioMarin.

Noella Noronha: no conflict of interest

Binu Abraham Philip: no conflict of interest

Omotola Olasupo: no conflict of interest

Drashti Pete: no conflict of interest

Tamara Navarro: no conflict of interest

Arun Keepanasseril: no conflict of interest

Davide Matino: has received research support from Bayer, Bioverativ/Sanofi, and Pfizer; and honoraria for speaking/participating in advisory boards from Bayer, Bioverativ/Sanofi, BIOVIIIx, Pfizer, and Sigilon.

Kerstin de Wit: no conflict of interest

Sameer Parpia: no conflict of interest

Alfonso Iorio's institution received research funding from NovoNordisk, Roche, Takeda, Bayer, Pfizer, BioMarin, CSL, Freeline, Grifols, Octapharma, Sanofi, Spark, and Uniqure.

## Authors' contributions

FG and AI conceived this review. FG is the guarantor of the review and drafted the manuscript. TN developed the search strategy. FG, NN, VBD, APB, and DP screened the articles and extracted the data. All authors contributed to the study design and interpretation of data. All authors read, provided feedback, and approved the final manuscript.

## References

1 Srivastava A, Santagostino E, Dougall A, *et al.* WFH Guidelines for the Management of Hemophilia, 3rd edition. *Haemophilia : the official journal of the World Federation of Hemophilia* 2020;**26**:1–158. doi:10.1111/hae.14046

2 Manco-Johnson MJ, Abshire TC, Shapiro AD, *et al.* Prophylaxis versus episodic treatment to prevent joint disease in boys with severe hemophilia. *The New England journal of medicine* 2007;**357**:535–44. doi:10.1056/NEJMoa067659

3 Feldman BM, Pai M, Rivard GE, *et al.* Tailored prophylaxis in severe hemophilia A: Interim results from the first 5 years of the Canadian Hemophilia Primary Prophylaxis Study. *Journal of Thrombosis and Haemostasis* 2006;**4**:1228–36. doi:10.1111/j.1538-7836.2006.01953.x

4 Hazendonk HCAM, Lock J, Mathôt RAA, *et al.* Perioperative treatment of hemophilia A patients: blood group O patients are at risk of bleeding complications. *Journal of Thrombosis and Haemostasis* 2016;**14**:468–78. doi:10.1111/jth.13242

5 Broderick CR, Herbert RD, Latimer J, *et al.* Association between physical activity and risk of bleeding in children with hemophilia. *JAMA - Journal of the American Medical Association* 2012;**308**:1452–9. doi:10.1001/jama.2012.12727

6 Abrantes JA, Solms A, Garmann D, *et al.* Relationship between factor VIII activity, bleeds and individual characteristics in severe hemophilia A patients. *Haematologica* 2020;**105**:1443–53. doi:10.3324/HAEMATOL.2019.217133

7 Collins PW, Blanchette VS, Fischer K, *et al.* Break-through bleeding in relation to predicted factor VIII levels in patients receiving prophylactic treatment for severe

hemophilia A. *Journal of thrombosis and haemostasis : JTH* 2009;**7**:413–20. doi:10.1111/j.1538-7836.2008.03270.x

8 Mahlangu J, Oldenburg J, Paz-Priel I, *et al.* Emicizumab Prophylaxis in Patients Who Have Hemophilia A without Inhibitors. *The New England journal of medicine* 2018;**379**:811–22. doi:10.1056/NEJMoa1803550

9 Rangarajan S, Walsh L, Lester W, *et al.* AAV5–Factor VIII Gene Transfer in Severe Hemophilia A. *New England Journal of Medicine* 2017;**377**:2519–30. doi:10.1056/NEJMoa1708483

10 George LA, Sullivan SK, Giermasz A, *et al.* Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. *The New England journal of medicine* 2017;**377**:2215–27. doi:10.1056/NEJMoa1708538

11 Moons KGM, de Groot JAH, Bouwmeester W, *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Medicine* 2014;**11**:e1001744. doi:10.1371/journal.pmed.1001744

12 Riley RD, Moons KGM, Snell KIE, *et al.* A guide to systematic review and meta-analysis of prognostic factor studies. BMJ (Online). 2019;**364**. doi:10.1136/bmj.k4597

13 Fischer K, van der Bom JG, Molho P, *et al.* Prophylactic versus on-demand treatment strategies for severe haemophilia: a comparison of costs and long-term outcome. *Haemophilia : the official journal of the World Federation of Hemophilia* 2002;**8**:745–52.

14 Andersson NG, Chalmers EA, Kenet G, *et al.* Mode of delivery in hemophilia: vaginal delivery and Cesarean section carry similar risks for intracranial hemorrhages and other major bleeds. *Haematologica* 2019;**104**:2100–6. doi:10.3324/haematol.2018.209619

15 Dolatkhah R, Khoshbaten M, Asvadi Kermani I, *et al.* Upper gastrointestinal bleedings in patients with hereditary coagulation disorders in Northwest of Iran: prevalence of Helicobacter pylori infection. *European journal of gastroenterology & hepatology* 2011;**23**:1172–7. doi:10.1097/MEG.0b013e32834b0e7a

16 Wong SS-L, Wilczynski NL, Haynes RB, *et al.* Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA . Annual Symposium proceedings AMIA Symposium* 2003;**2003**:728–32.

17 Ouzzani M, Hammady H, Fedorowicz Z, *et al.* Rayyan-a web and mobile app for systematic reviews. *Systematic reviews* 2016;**5**:210. doi:10.1186/s13643-016-0384-4

18 Hayden JA, van der Windt DA, Cartwright JL, *et al.* Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013;**158**:280–6. doi:10.7326/0003-4819-158-4-201302190-00009

19 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine* 2009;**6**:e1000097. doi:10.1371/journal.pmed.1000097

20 Ahnström J, Berntorp E, Lindvall K, *et al.* A 6-year follow-up of dosing, coagulation factor levels and bleedings in relation to joint status in the prophylactic treatment of haemophilia. Haemophilia. 2004;**10**:689–97. doi:10.1111/j.1365-2516.2004.01036.x

21 Fosser C, Roberts J, Tortorici M, *et al.* Identifying Efficacious Thresholds for Bleeding Risk Reduction in Relation to Factor IX (FIX) Levels in Hemophilia B Patients Receiving IDELVION. In: *Journal of Pharmacokinetics and Pharmacodynamics*. 2017. S15–S15.

22 Dargaud Y, Negrier C, Rusen L, *et al.* Individual thrombin generation and spontaneous bleeding rate during personalized prophylaxis with Nuwiq® (human-cl rhFVIII)

in previously treated patients with severe haemophilia A. *Haemophilia* 2018;**24**:619–27. doi:10.1111/hae.13493

23 Ross C, Goldenberg NA, Hund D, *et al.* Athletic participation in severe hemophilia: Bleeding and joint outcomes in children on prophylaxis. *Pediatrics* 2009;**124**:1267–72. doi:10.1542/peds.2009-0072

24 Anderson A, Forsyth A. Playing it safe: Bleeding disorders, sports and exercise. *New York, NY: National Hemophilia Foundation* 2005;**44**.

25 Desjonqueres A, Guillet B, Beurrier P, *et al.* Bleeding risk for patients with haemophilia under antithrombotic therapy. Results of the French multicentric study ERHEA. British Journal of Haematology. 2019;**185**:764–7. doi:10.1111/bjh.15606

26 Gupta S, Siddiqi AEA, Soucie JM, *et al.* The effect of secondary prophylaxis versus episodic treatment on the range of motion of target joints in patients with haemophilia. *British Journal of Haematology* 2013;**161**:424–33. doi:10.1111/bjh.12267

27 Jobe SM, Dunn AL, Leong T. Procoagulant Platelet Potential Is Inversely Correlated with Bleeding and Joint Disease in Severe Hemophilia A. *Blood* 2018;**132**:1185.

28 Fischer K, Collins P, Björkman S, *et al.* Trends in bleeding patterns during prophylaxis for severe haemophilia: Observations from a series of prospective clinical trials. *Haemophilia* 2011;**17**:433–8. doi:10.1111/j.1365-2516.2010.02450.x

29 Ainsworth BE. How do i measure physical activity in my patients? Questionnaires and objective methods. British Journal of Sports Medicine. 2009;**43**:6–9. doi:10.1136/bjsm.2008.052449

30 Rütten A, Ziemainz H, Schena F, *et al.* Using different physical activity measurements in eight European countries. Results of the European Physical Activity Surveillance System

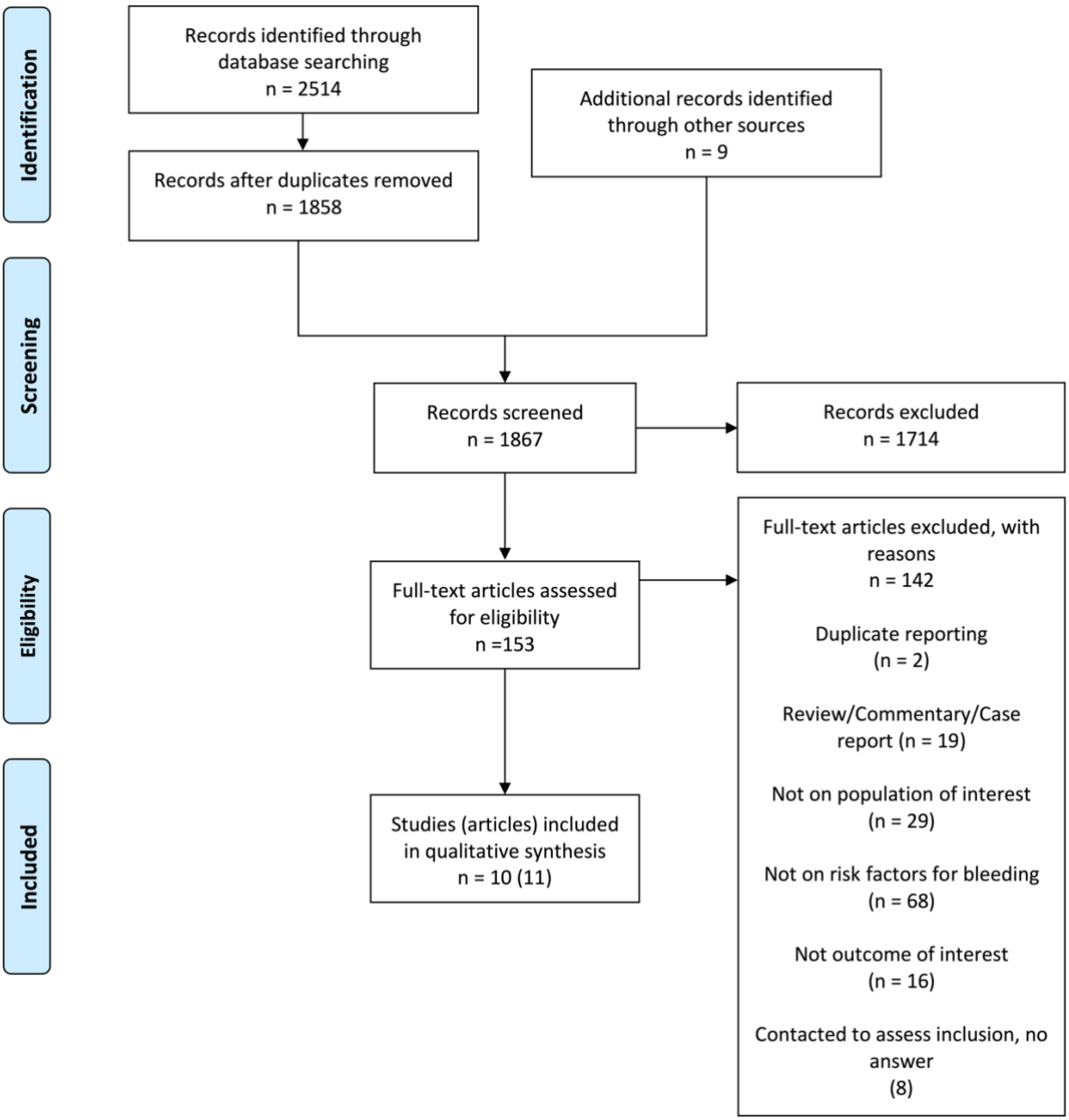
(EUPASS) time series survey. *Public Health Nutrition* 2003;**6**:371–6. doi:10.1079/phn2002450

31 Santagostino E, Martinowitz U, Lissitchkov T, *et al.* Long-acting recombinant coagulation factor IX albumin fusion protein (rIX-FP) in hemophilia B: Results of a phase 3 trial. *Blood* 2016;**127**:1761–9. doi:10.1182/blood-2015-09-669234

32 Pettersson H, Gilbert MS. Hemophilic arthropathy. In: *Diagnostic imaging in hemophilia*. Springer 1985. 23–68.

## Figures

*Figure 2-4: PRISMA Flow Diagram.*




From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097
For more information, visit www.prisma-statement.org.

## Tables

*Table 2-9: characteristics of the included studies.*

| Author, year | Study design | Country | Type of hemophilia | Hemophilia severity[$] | Treatment (dose) | Median age (Q1; Q3) | Participants (n), (percentage of severe) | Events (n) | FUP duration (months) |
|---|---|---|---|---|---|---|---|---|---|
| Abrantes, 2020[6] | Interventional | North America, Argentina, Colombia, Europe, Asia. | A | Severe | Regular prophylaxis (20-50 IU/kg, 2-3x/week) | 22 (1, 61) | 172 (100%) | 633 | 12 |
| Ahnstrom, 2004[20] | Cohort, retrospective | Sweden | A; B | Moderate; Severe | Regular prophylaxis (12-68 IU/kg, 0.7-7x/week) | NA | 64 (97%) | NA | 72 |
| Broderick, 2012[5] | Case-crossover, prospective | Australia | A; B | Moderate; Severe | Prophylaxis; on-demand (99-107 IU/kg weekly)[#] | 10 (4)* | 104 (83%) | 329 | 12 |
| Collins, 2009[7] | Interventional | USA, Canada, Europe | A | Severe | Regular prophylaxis (84-108 IU/kg weekly)~ | NA | 143 (100%) | 924 | 12 |
| Dargaud, 2018[22] | Cohort, prospective | Europe | A | Severe | Regular prophylaxis (95 IU/kg weekly) | 33 (9)* | 32 (100%) | 7 | 6 |
| Desjonqueres, 2019[25] | Case-control, prospective | France | A; B | All | Prophylaxis (NR); on-demand | 62 (50; 94)^ | 126 (10%[@]) | NR | NR |
| Fischer, 2011[28]" | Interventional | USA, Canada, Europe | A | Severe | Regular prophylaxis | NA | 145 (100%) | 327 | 12 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | (84-108 IU/kg weekly)~ | | | | |
| Fosser, 2017[21,31] | Interventional | North America, Europe, Asia, Middle East | B | Moderate; Severe (FIX <2 IU/dL) | Prophylaxis (50 IU/kg weekly or 75 IU/kg every 2 weeks-10 days); on-demand | 12-61[%] | 57 (87%) | 478 | 12 |
| Gupta, 2013[26] | Cohort, retrospective | USA | A; B | Moderate; Severe | Prophylaxis (NR); on-demand | 22 (13)* | 286 (78%) | NR | 24 |
| Jobe, 2018[27] | Cohort, retrospective | USA | A | Severe | Prophylaxis (NR); on-demand | NR | 92 (100%) | NR | NR |
| Ross, 2009[23] | Cohort, retrospective | USA | A; B | Moderate; Severe (FVII or FIX <2 IU/dL) | Regular (92%) or situational prophylaxis (NR)[&] | 5-20[%] | 37 (97%) | NR | 12 |

"same population than Collins 2009

[$]Unless otherwise specified, severe hemophilia was defined as factor VIII or IX levels <1 IU/dL, and moderate hemophilia as 1 IU/dL ≤ factor VIII or IX levels <1 IU/dL

[#]median weekly dose for people with hemophilia B and A, respectively

[~]median weekly dose for people aged 10-65 and 1-6 years, respectively

[&]95% of participants infused at least twice weekly

[+]not reported if prospective

*mean (standard deviation)

^median (min; max)

[%]range

[@]of the 44 cases, for a total of 4 severe patients

FVIII: factor VIII; FIX: factor IX; FUP: follow up; NR: not reported; UK: United Kingdom; USA: United States of America

*Table 2-10: Risk of bias assessment of the included studies, using the QUIPS tool.*

| Author, year | Selection bias - Study Participation | Study Attrition | Prognostic Factor Measurement | Outcome Measurement | Adjustment for other predictors* | Stat Analysis |
|---|---|---|---|---|---|---|
| Abrantes, 2020[6] | High | High | Low | Low | High | Low |
| Ahnstrom, 2004[20] | Low | Probably low | High (factor levels)<br>Low (joint score) | Probably low | High | Low |
| Broderick, 2012[5] | Low | Probably low | Probably low | Low | Low | Low |
| Collins, 2009[7] | High | Probably low | Probably low | Low | High | Low |
| Dargaud, 2018[22] | High | Low | Low | Low | High | Low |
| Desjonqueres, 2019[25] | High | Low | Probably low | Probably low | High | High |
| Fischer, 2011[28] | High | Probably low | Low | Low | High | Low |
| Fosser, 2018[21,31] | High | Probably low | High | Low | High | Probably high |
| Gupta, 2013[26] | High | High | Low | High | High | Low |
| Jobe, 2018[27] | Probably low | Probably low | Low | Probably high | High | High |
| Ross, 2009[23] | High | High | High (Physical activity)<br>Low (age)<br>Probably high (Injuries)<br>Probably low (infusions) | Probably low | High | High |

*namely “Study confounding” in the original version of the tool.

*Table 2-11: risk factors, outcomes definitions, risk estimates, and covariates adjusted for.*

| Prognostic factor | Definition/measurement method | Author, year | Outcome | Definition | Risk estimate | Adjustment |
|---|---|---|---|---|---|---|
| Factor levels/treatment pattern | | | | | | |
| Plasma factor levels | Based on individual PK and treatment diary | Abrantes, 2020[6] | All bleeds | All self-reported bleeds | FVIII levels constantly at 5 IU/dL would reduce the ABR by ~1/3 | Age, body weight, BMI, lean body weight, race, vWF, treatment, n of target joints. |
| | Based on PK (full or pop), treatment regimens (ITT) | Ahnstrom, 2004[20] | Joint bleeds; non-joint bleeds | Joint pain + stiffness and swelling | NR (weak association for joint bleeds, no association for other bleeds) | Joint score (stratified analysis)[32] |
| | Based on full PK and last factor infusion (not clear if based on diary, interview, or ITT). | Broderick, 2012[5] | All bleeds | Treated bleeds | aOR 0.98 (0.97; 0.99) | Plasma factor levels, physical activity, and individual characteristics (through the case-cross over design) |
| | Hours spent < 1 IU/dL /week | Collins, 2009[7] | All bleeds | Treated bleeds | Annual bleeding rate increased by 2.2% (1.58; 2.78) in PWH aged 1-6 years and 1.4% (0.21; 2.62) in 10-65 years | Bleed cause, bleed site, age, and weight ratio (actual weight/ideal weight) |
| | At baseline | Dargaud, 2018[22] | Spontaneous bleeds | Self-reported non- | AUROC 0.58 | None |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | traumatic non-operative bleeds | (0.33; 0.84) | |
| | Through, based on PopPK, no more details | Fosser, 2018[21,31] | All bleeds | Treated bleed | Daily risk for bleeding reduction FIX trough level >2 IU/mL: 69% (53; 80) >5: 77% (67; 84) >10: 78% (69; 85) | None |
| Prophylaxis frequency | Number of factor infusions per week | Ross, 2009[23] | Frequent joint bleeds | ≥1 bleed in a joint/ ~8 weeks | OR 1.07 (0.42; 2.73) per infusion | None |
| Adherence to treatment frequency | Percentage of weeks when treated on ≥3 days | Collins, 2009[7] | All bleeds | Treated bleeds | 0.97 (95% CI 0.63; 1.27) fewer bleeds/year in the 1–6 years old; 1.19 (CI 0.66–1.61) fewer in the 10–65 | Bleed cause, bleed site, age, and weight ratio (actual weight/ideal weight) |
| Physical activity levels | Self-reported, according to the ANHF.[24] | Broderick, 2012[5] | All bleeds | Treated bleeds | Significant collisions possible: aOR 2.7 (95%CI 1.7; 4.8) Significant collisions inevitable: aOR 3.7 (2.3; 7.3) Ref: non-contact activities | Plasma factor levels, physical activity, and individual characteristics (through the case-cross over design) |
| | High versus low impact, simplified from WFH definitions | Ross, 2009[23] | Frequent joint bleeds | ≥1 bleed in a joint/ ~8 weeks | aOR 0.32 (0.04; 2.70) | Prophylaxis frequency |

| Previous injuries/bleeding history | | | | | | |
|---|---|---|---|---|---|---|
| Injuries per season (n) | NR | Ross, 2009[23] | Frequent joint bleeds | ≥1 bleed in a joint/ ~8 weeks | OR 7.02 (0.30; 167) per increase of one | None |
| n of bleeds in the previous 12 months | Based on diaries | Abrantes, 2020[6] | All bleeds | All self-reported bleeds | 1 bleed: hazard 54% lower (95%CI 40-65) 84 bleeds: hazard 147% higher (95%CI 79-226) Ref: mean patient with 8.2 bleeds | Age, body weight, BMI, lean body weight, race, vWF, treatment, n of target joints. |
| Previous non-severe bleeds | Not requiring hospitalization or any treatment | Desjonquer es, 2019[25] | Severe bleeds | Requiring hospitalizati on or any treatment | OR 21 (p 0.001) | None |
| Frequent non-serious bleeds | > 1/month | Desjonquer es, 2019[25] | Severe bleeds | Requiring hospitalizati on or any treatment | OR 169 (p 0.038) | None |
| Age at target joint developmen t | <5 y | Gupta, 2013[26] | Joint bleeds | Self-reported bleeding located in a joint | ref | Age at target developme nt, BMI, hemophilia severity, treatment type (prophylaxis vs on-demand) presence of inhibitors. |
| | 5-9 y | | | | aIRR 2.78 (2.07; 3.74) | |
| | 10-14 y | | | | aIRR 2.93 (2.19; 3.93) | |
| | 15-19 y | | | | aIRR 2.74 (2.04; 3.68) | |
| | ≥20 y | | | | aIRR 2.72 (2.04; 3.64) | |
| Age | In years | Ross, 2009[23] | Frequent joint bleeds | ≥1 bleed in a joint/~8 weeks | OR 1.04 (0.81; 1.32) per year | None |
| BMI | Underweight/norma l | Gupta, 2013[26] | Joint bleeds | Self-reported bleeding | Ref | Age at target developme |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Overweight | | | located in a joint | aIRR 1.05 (95% CI 0.98; 1.13 | nt, BMI, hemophilia severity, treatment type (prophylaxis vs on-demand) presence of inhibitors. |
| | Obesity | | | | aIRR: 1.11 (1.04; 1.20) | |
| Antithrombotic therapy | antiplatelets or anticoagulants | Desjonqueres, 2019[25] | All bleeds | All self-reported bleeds | NR (no association) | None |
| | antiplatelets or anticoagulants | Desjonqueres, 2019[25] | Severe bleeds | Requiring hospitalization or any treatment | OR 3.55 (1.2; 10.4) | None |
| Blood tests | | | | | | |
| Endogenous thrombin potential | Thrombin generation assay (TGA) | Dargaud, 2018[22] | Spontaneous bleeds | Self-reported non-traumatic non-operative bleeds | AUROC 0.94 (0.88; 1.00) | None |
| Procoagulant platelet potential | Ratio of procoagulant platelets and all activated platelets following stimulation with thrombin and convulxin | Jobe, 2018[27] | All bleeds | NA | correlation coefficient with ABR r = -0.47, p < 0.0001 | None |
| P-selectin expression | Measured on washed platelets | Jobe, 2018[27] | All bleeds | NA | NR (not statistically significant) | None |
| procaspase-activating compound 1 (PAC-1) | Measured on washed platelets | Jobe, 2018[27] | All bleeds | NA | NR (not statistically significant) | None |
| Season of the year | Summer: June–August Autumn: September–November Winter: December–February Spring: March–May | Fischer, 2011[28] | Joint bleeds | Treated bleed, located in a joint | 1-6 y: no seasonal variation 10–17 y: 43% all joint bleeds in summer 18–65 y: 46% all joint | Adherence to frequency, ratio weight (actual weight divided by ideal weight for height), age, and time below |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | bleeds in summer | 1% FVIII level |

ABR: annualized bleeding rate; ANHF: American National Hemophilia Foundation; AUROC: area under the receiver operating characteristic curve; BMI: body mass index; ITT: intention to treat; NR: not reported; OR: Odds ratio; PK: pharmacokinetics; vWF: von Willebrand factor; WFH: world federation of hemophilia.

# Chapter 3 – Measuring physical activity

Germini, F, MD[1,2]; Noronha, N, MSc[1]; Borg Debono V, PhD[1]; Philip, B.A., MSc[1], Pete, D, MPH[1]; Navarro, T, MLIS, MEd[1]; Keepanasseril, A, MSc[1]; Parpia, S, PhD[1, 3] de Wit, K, MD[1,2,4] and Iorio, A, MD, PhD[1,]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Oncology, McMaster University, Hamilton, ON, Canada

[4]Department of Emergency Medicine, Queen's University, Kingston, ON, Canada

## Accuracy and acceptability of wrist-wearable activity tracking devices: a systematic review of the literature.

### Abstract

*Background and objectives:* numerous wrist-wearable devices to measure physical activity are currently available, but there is a need to unify the evidence about how they compare in terms of acceptability and accuracy. We performed a systematic review of the literature to assess the accuracy and acceptability (willingness to use the device for the task it is designed to support) of wrist-wearable activity trackers.

*Methods:* we searched MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials (CENTRAL) and SPORTDiscus for studies measuring physical activity in the general population,

using wrist-wearable activity trackers. We screened articles for inclusion and, for included studies, reported data on the studies' setting and population, outcome measured, and risk of bias.

*Results:* 65 articles were included in our review. Accuracy was assessed for 14 different outcomes, that can be classified in the following categories: count of specific activities (including step counts), time spent being active, intensity of physical activity (including energy expenditure), heart rate, distance, and speed. Substantial clinical heterogeneity did not allow to perform a meta-analysis of the results. The outcomes assessed more frequently were step counts, heart rate, and energy expenditure. For step counts, Fitbit Charge (or Charge HR) had a mean absolute percent error (MAPE) < 25% across 20 studies. For heart rate, Apple watch had a MAPE < 10% in 2 studies. For energy expenditure, the MAPE > 30% for all the brands, showing poor accuracy across devices. Acceptability was more frequently measured through data availability and wearing time. Data availability was $\geq$ 75% for FitBit Charge HR, FitBit Flex 2, and Garmin Vivofit. The wearing time was 89% for both GENE Activ and Nike Fuelband.

*Conclusions:* Fitbit Charge and Charge HR were consistently shown to have a good accuracy for step counts and Apple watch for measuring heart rate. None of the tested devices proved to be accurate in measuring energy expenditure. Efforts should be made to reduce the heterogeneity between studies.

*Key Words:* Diagnosis; measurement; wrist-wearable devices

## Background

Tracking, measuring and documenting one's physical activity can be a way of monitoring and encouraging a person to participate in daily physical activity; increased activity that is thought to translate into important, positive health outcomes, both physically and mentally.[1] In the past, most physical activity tracking was done manually by oneself or an external assessor, through records, logbooks, or using questionnaires. These are indirect methods to quantify physical activity, meaning that they do not measure movement directly as it occurs.[2] The main disadvantages of such methods are the administrative burden on either the self-assessor, or the external assessor, and the potential imprecision due to recall bias.[2,3] Direct methods to assess physical activity,[2] such as accelerometer or pedometers that digitally record movement are preferred, because they eliminate recall bias and are convenient. This process of activity tracking has become automated, accessible and digitized with wearable tracking technology such as wristband sensors and smartwatches (that can be linked to computer applications on other devices such as smartphones, tablets and personal computers). When data are uploaded to these latter devices, a person can then review their physical activity log and potentially use this feedback to make behavioral changes to physical activity.

The ideal device should be acceptable to the end-user, affordable, easy to use, and accurate in measuring physical activity. Accuracy can be defined as the closeness of the measured value to the actual value. Accuracy can be calculated using measures of agreement, sensitivity and specificity, receiver operating characteristic curves (ROCs), or absolute and percentage differences.[4] Agreement can be defined as "the degree of concordance between two or more sets of measurements".[5] It can be measured as percentage agreement, i.e. the percentage of cases in which two methods of measurements of the same variable lead to classification in the same

category. Another example of methods of calculating agreement is the kappa statistics, that measures agreement beyond chance.[6] Sensitivity and specificity are the true positive and true negative proportion, respectively. These proportions are calculated using the measurement method that we are evaluating as the index test, and another method, known to be accurate, as the reference standard.[7] ROC curves are obtained plotting the sensitivity vs the complement of specific and can be used to find optimal cut-off points for the index test. Absolute and percentage differences are used to determine how far the index test measurement is from the reference standard, or their average.[7] Acceptability can be widely defined as "the demonstrable willingness within a user group to employ information technology for the task it is designed to support".[8] It can be assessed qualitatively (e.g. through questionnaires or interviews) or quantitatively (e.g. percentage of the time in which the device is worn or the data are available, or measured using ad hoc scales). Based on a 2019 review, acceptability or acceptance of wrist-wearable activity tracking devices is dependent on the type of user and context of use.[9] This same review indicates that research on accuracy has not kept up with the plethora of wearable physical activity tracking devices in the market.[9] This may be because of the rapidly changing landscape as companies continue to upgrade models with different technical specifications and features. The purpose of this systematic review is to assess the acceptability and accuracy of these wrist-wearable activity tracking devices through a focused in-depth review of primary studies assessing these two characteristics.

**Objectives**

The first objective of this systematic review was to assess accuracy of wrist-wearable activity tracking devices for measuring physical activity.

The second objective was to assess the acceptability of wrist-wearable activity tracking devices for measuring physical activity.

## Methods

The methods for this systematic review have been registered in the PROSPERO database (CRD42019137420).

### Search strategy

Databases searched include MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials (CENTRAL) and SPORTDiscus, from inception to May 28, 2019. Search strategies were developed to retrieve content on wearable activity tracker and on their accuracy and reproducibility of results. We used search terms including *Wearable device*, and *Fitness Tracker* to identity studies on the use of a consumer-based wearable activity tracker; while terms such as *data accuracy and reproducibility of results* were included to bring in content focused on tracker activity validation. The search strategy is available online in the PROSPERO record. A snowball search was conducted by checking the references of relevant studies and systematic reviews on this topic that were identified in our original search.

### Selection of studies

For the acceptability objective, the population was the general population, without sex or age restrictions. The intervention was the use of a wrist-wearable activity tracker. The outcome was any quantitative measure of acceptability, including wearing time, data availability, or questionnaires to assess the acceptability.

For the accuracy objective, the population was the same as above, the index test had to be a wrist-wearable activity tracker, and the reference standard could be another device or any method to measure physical activity, including questionnaires or direct observation. The outcome could be any measure of physical activity, including but not limited to step count, heart rate, distance, speed, activity count, activity time, or intensity of physical activity.

For both the objectives, this review examined both research grade devices (activity trackers available only for research purposes) and commercial devices (those available to the general public). We included studies were limited to community-based everyday life setting. Laboratory tests such as research studies were included as long as everyday settings were reproduced – therefore, excluding institutionalized patients and hospitalized situations. We set no restrictions on the length of observation for the original studies.

Exclusion criteria: device not worn on the wrist, studies measuring sleep, studies on institutionalized or hospitalized patients.

Study characteristics: All studies reporting primary data were considered for inclusion, with the exception of case reports and case series.

Using these pre-specified inclusion and exclusion criteria and a piloted form, we initially screened for inclusion from the titles and abstracts of the retrieved articles, using the online software Rayyan.[10] Subsequently, we screened the full texts of the studies identified as potentially eligible from the title and abstract screening for selection.

**Data extraction and risk of bias assessment**

Data were extracted on an Excel file. The data extraction form was based on a previous publication on the same topic[9] and adapted to the needs of this review. The following data were extracted: general study information: first author's name, publication year, type of study (perspective vs retrospective, observational vs interventional), duration of follow up (in days), setting (laboratory versus field); characteristics of the population: number of participants, underlying disease (e.g. healthy subjects, people with severe obesity, chronic joint pain…), gender, and age distribution [mean and standard deviation (SD), or median and min-max or first and third quartile); measures of accuracy: step count, distance, speed, heart rate, activity count, time spent

being active, intensity of physical activity; and acceptability of the device including, but not limited to, data availability, wearing time, ease of use. The risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies – version 2 (QUADAS-2) tool.[11] This tool guides the assessment of the risk for bias in diagnostic accuracy studies in four domains: patient selection, index test, reference standard, and flow and timing. We rated the risk for bias in each domain as "High", "Probably high", "Probably low", and "Low". When necessary, the study authors were contacted for additional information.

Throughout title and abstract and full text screening and the data extraction, each step was performed in duplicate with two reviewers deciding independently on inclusion or exclusion, and if needed, later discussing with another author to make a final decision. Disagreements were solved through discussion and, when needed, with the intervention of a third reviewer. The reviewers were trained with calibration exercises until an adequate performance was achieved for each of these steps.

**Diagnostic accuracy measures**

When available, we extracted the mean absolute percentage error (MAPE) or the mean percentage error. When these were not available, we extracted other measures, in the following order of priority: mean difference, mean bias (Bland-Altman), accuracy determined through intraclass correlation coefficient (ICC), and correlation coefficient (Pearson's or Spearman's). When the outcome was dichotomized and sensitivity and specificity were calculated, we reported on those. When available, we reported measures of variability or 95% confidence intervals (CI) for all the above-mentioned measures. The formula for calculating the MAPE, mean percentage error, mean difference, and mean bias are reported in the *supplementary material.*

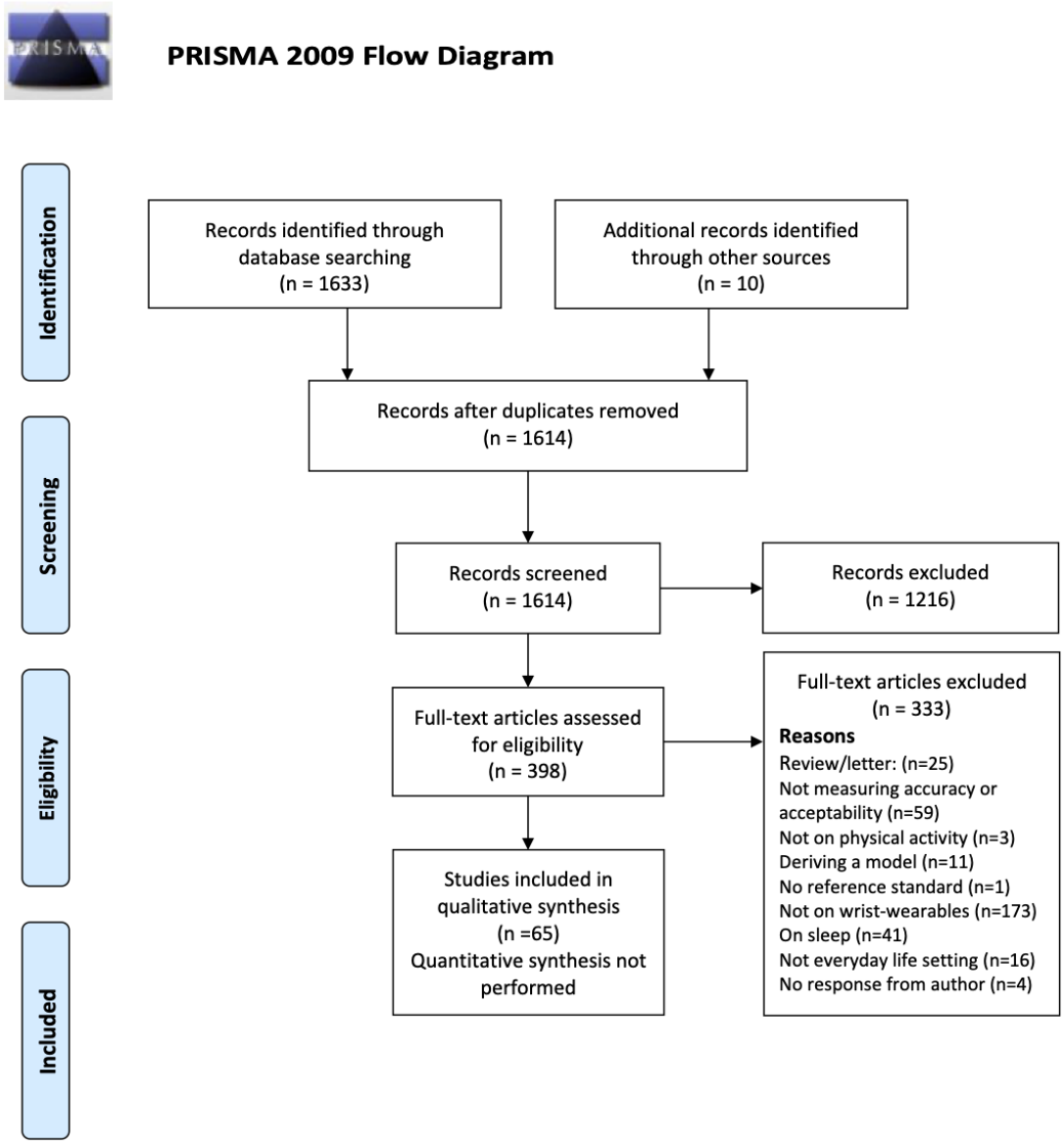
**Synthesis of results**

Due to the significant heterogeneity observed in the studies' population, setting, devices assessed, reference standard, outcomes assessed, and the outcome measures reported, we decided not to perform a quantitative synthesis, and provided a narrative synthesis of the results for both the objectives. For the accuracy objective, given the high number of studies retrieved, we summarized only on devices that were included in at least two studies reporting the same outcome. All the remaining results were reported in the supplementary material.

## Results

The search identified 1633 records (1614 after duplicates removal). The study flow diagram is reported in Figure 2-1. After screening the full text of 398 articles, 65 articles have been included in the systematic review. The characteristics of the included studies are summarized in Table 2-1 and Table 2-2 for the accuracy and acceptability objective, respectively. All the included studies were single center, with a prospective, observational design. The complete results for both the accuracy and acceptability objectives are reported in the *supplementary material.*

Figure 3-1. PRISMA Study Flow Diagram.




*From:* Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

**For more information, visit www.prisma-statement.org.**

*Table 3-1:12 Characteristics of the studies reporting on accuracy.*

| Author, year | Setting | FUP time | Sample | Mean age (SD) | Female % | Underling disease | Outcome | Device brand and model |
|---|---|---|---|---|---|---|---|---|
| Alharbi, 2016[12] | Lab | < 1 day | 48 | 66 (7) | 48 | Cardiac rehabilitation patients | Step count<br>MVPA | Fitbit Flex |
| Alsubheen, 2016[13] | Field | 5 days | 13 | 40 (12) | 38 | Healthy | Step count<br>Energy expenditure | Garmin Vivofit |
| An, 2017[14] | Lab and Field | 1 day | 35 | 31 (12) | 51 | Healthy | Step count | Fitbit Flex; Garmin Vivofit; Polar Loop; Basis B1 Band; Misfit Shine; Jawbone UP24; Nike FuelBand SE |
| An, 2017[15] | Field | < 1 day | 62 | 24 (5) | 40 | Healthy | Active time | ActiGraph GT3X |
| Blondeel, 2018[16] | Field | 14 days | 8 | 65 (8) | 25 | COPD | Step count | Fitbit Alta |
| Boeselt, 2016[17] | Field | 3 days | 20 | 66 (7) | 15 | COPD | Step count<br>Energy expenditure<br>MVPA | Polar A300 |
| Bruder, 2018[18] | Lab and Field | 395 days | 32 | NA | NA | Post radial fracture rehab | Activity count | ActivPAL |
| Bulathsinghala, 2014[19] | Lab | 1 day | 20 | 70 (10) | NA | COPD | Physical activity intensity | ActiGraph GT3X+ |
| Burton, 2018[20] | Lab | < 1 day | 31 | 74 (6) | 65 | Healthy elderly | Step count | Fitbit Flex; Fitbit Charge HR |
| Choi, 2010[21] | Lab | < 1 day | 76 | 13 (2) | 62 | Healthy | Energy expenditure | ActiGraph GT1M |
| Chow, 2017[22] | Lab | 1.5 days | 31 | 24 (5) | 39 | Healthy | Step count | ActiGraph wGT3xBT-BT; Fitbit Flex; Fitbit Charge HR; Jawbone UP24 |
| Chowdhury, 2017[23] | Lab and Field | 2 days | 30 | 27 (6) | 50 | Healthy | Energy expenditure | Microsoft Band; Apple Watch, series not NA; Jawbone Up24; Fitbit Charge |
| Cohen, 2010[24] | Lab and Field | 3 days | 57 | 70 (10) | NA | COPD | Speed | ActiGraph Mini MotionLogger |
| Compagnat, 2018[25] | Lab | < 1 day | 46 | 65 (13) | NA | Stroke | Energy expenditure | ActiGraph GT3X+ |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dondzila, 2018[26] | Lab and Field | 3-62 days | 40 | 22 (2) | 58 | Healthy | Step count<br>Energy expenditure<br>Heart rate | Fitbit Charge HR; Miio FUSE |
| Dooley, 2017[27] | Lab | 1 day | 62 | 23 (4) | 58 | Healthy | Heart rate<br>Energy expenditure | Apple Watch, series not NA; Fitbit Charge HR; Garmin Forerunner 225 |
| Durkalec-Michalski, 2013[28] | Lab | 2 days | 20 | 26 (5) | 55 | Healthy | Energy expenditure | ActiGraph GT1M |
| Falgoust , 2018[29] | Lab | 1 day | 30 | NA | NA | Healthy | Step count | Fitbit Charge HR; Fitbit Surge; Garmin Vivoactive HR |
| Ferguson, 2015[30] | Field | 2 days | 21 | 33 (10) | 52 | Healthy | Step count<br>MVPA<br>Energy expenditure | Nike Fuelband; Misfit Shine; Jawbone UP |
| Gaz, 2018[31] | Lab and Field | < 1 day | 32 | 36 (8) | 69 | Healthy | Step count<br>Distance | Fitbit Charge HR; Apple Watch, series not NA; Garmin Vivofit 2; Jawbone UP2 |
| Gillinov, 2017[32] | Lab | < 1 day | 50 | 38 (12) | 54 | Healthy | Heart rate | Garmin Forerunner 235; TomTom Spark; Apple Watch, series not NA; Fitbit Blaze |
| Gironda, 2007[33] | Lab | < 1 day | 3 | 43 | 31 | Pain syndromes | Activity count | Actiwatch Score |
| Hargens, 2017[34] | Lab and Field | 7 days | 21 | 31 | 68 | Healthy | MVPA<br>Energy expenditure<br>Step count | Fitbit Charge |
| Hernandez-Vicente, 2016[35] | Field | 7 days | 18 | 21 (1) | 50 | Healthy | Energy expenditure<br>Vigorous active time<br>Active time<br>Step count | Polar V800 |
| Huang, 2016[36] | Lab | 1 day | 40 | 24 (3) | 25 | Healthy | Step count<br>Distance | Jawbone UP24; Garmin Vivofit; Fitbit Flex; Nike Fuelband |
| Imboden, 2018[37] | Lab | < 1 day | 30 | 49 (19) | 50 | Healthy | Energy expenditure<br>MVPA<br>Step count | Fitbit Flex; Jawbone Up24; Fitbit Flex |
| Jo, 2016[38] | Lab | < 1 day | 24 | 25 | 50 | Healthy | Heart rate | Basis Peak K; Fitbit Charge |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Jones, 2018[39] | Lab | 118 days | 30 | 33 | NA | Healthy | Step count | Fitbit Flex |
| Kaewkannate, 2016[40] | Lab | < 1 day | 7 | 31 (0) | 14 | Healthy | Step count | Fitbit Flex; Jawbone UP24; Withings Pulse; Misfit Shine |
| Lamont, 2018[41] | Lab | < 1 day | 33 | 67 (8) | 64 | Parkinson's | Step count | Garmin Vivosmart HR; Fitbit Charge HR |
| Lauritzen, 2013[42] | Lab and Field | < 1 day | 18 | NA | 56 | Elderly | Step count | Fitbit Ultra |
| Lawinger, 2015[43] | Lab | < 1 day | 30 | 26 (6) | 70 | Healthy | Activity count | ActiGraph GT3X+ |
| Lemmens, 2018[44] | Lab | < 1 day | 40 | 31 (5) | 100 | Parkinson's | Energy expenditure | Phillips Optical Heart Rate monitor |
| Magistro, 2018[45] | Lab | < 1 day | 40 | 74 (7) | 60 | Healthy | Step count | ADAMO Care Watch |
| Mandigout, 2017[46] | Lab | < 1 day | 24 | 68 (14) | 60 | Stroke | Energy expenditure | Actical; ActiGraph GTX |
| Manning, 2016[47] | Lab | < 1 day | 9 | 15 (1) | NA | Severe obesity | Step count | Fitbit One; Fitbit Flex; Fitbit Zip; ActiGraph GT3x+; Jawbone UP |
| Montoye, 2017[48] | Lab | < 1 day | 30 | 24 (1) | 47 | Healthy | Step count<br>Energy expenditure<br>Heart rate | Fitbit Charge HR |
| Powierza, 2017[49] | Field | < 1 day | 22 | 22 (2) | 55 | Healthy | Heart rate | Fitbit Charge |
| Price, 2017[50] | Lab | < 1 day | 14 | 23 | 21 | Healthy | Energy expenditure | Fitbit One; Garmin Vivofit; Jawbone UP |
| Redenius, 2019[51] | Lab | 4 days | 65 | 42 (12) | 72 | Healthy | MVPA | Fitbit Flex |
| Reid, 2017[52] | Field | 4 days | 22 | 21 (2) | 100 | Healthy | MVPA<br>Step count | Fitbit Flex |
| Roos, 2017[53] | Lab | 2 days | 20 | 24 (2) | 40 | Runners | Energy expenditure | Suunto Ambi; Garmin Forerunner 920XT; Polar V800 |
| Schaffer, 2017[54] | Lab | < 1 day | 24 | 54 (13) | 42 | Stroke | Step count | Garmin Vivofit |
| Scott, 2017[55] | Field | 7 days | 89 | NA | 54 | Healthy | Daily mean activity<br>MVPA | GENE Activ |
| Semanik, 2020[56] | Lab | 7 days | 35 | 52 | 69 | Chronic joint pain | MVPA | Fitbit Flex |
| Sirard, 2017[57] | Lab and Field | 7 days | 14 | 9 (2) | 50 | Healthy | Energy expenditure<br>MVPA<br>Step count | Movband; Sqord |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| St-Laurent, 2018[58] | Lab | 7 days | 16 | 33 (4) | 100 | Pregnant | Step count MVPA | Fitbit Flex |
| Stackpool, 2013[59] | Lab | < 1 day | 20 | 22 (1) | 50 | Healthy | Step count Energy expenditure | Jawbone UP; Nike Fuelband; Fitbit Ultra; Adidas MiCoach |
| Stiles, 2013[60] | Lab | 1 day | 10* | 39 (6) | 100 | Healthy premenopausal women | Loading rate (BW/s) | GENE Activ; ActiGraph GT3X+ |
| Støve, 2019[61] | Lab | < 1 day | 29 | 29 (9) | 41 | Healthy | Heart rate | Garmin Forerunner |
| Tam, 2018[62] | Lab | < 1 day | 30 | 32 (9) | 50 | Healthy | Step count | Fitbit Charge HR; Miio Mi Band 2 |
| Thomson, 2019[63] | Lab | < 1 day | 30 | 24 (3) | 50 | Healthy | Heart rate | Apple Watch, series not NA; Fitbit Charge HR 2 |
| Wahl, 2017[64] | Lab | < 1 day | 20 | 25 (3) | 50 | Healthy | Step count Energy expenditure Distance | Polar Loop; Beurer AS80; Fitbit Charge HR; Fitbit Charge; Bodymedia Sensewear; Garmin Vivofit; Garmin Vivosmart; Garmin Vivoactive; Garmin Forerunner 920XT; Xaomi Mi Band; Withings Pulse |
| Wallen, 2016[65] | Lab | < 1 day | 22 | 24 (6) | 50 | Healthy | Heart rate Energy expenditure Step count | Apple Watch, series not NA; Samsung Gear S; Miio Mio alpha; Fitbit Charge |
| Wang, 2017[66] | Lab | < 1 day | 9 | 22 (1) | 44 | Healthy | Step count | Huawei B1; Mi Band Miband; Fitbit Charge; Polar Loop; Garmin Vivofit 2; Misfit Shine; Jawbone UP |
| Woodman, 2017[67] | Lab | < 1 day | 28 | 25 (4) | 29 | Healthy | Energy expenditure | Garmin Vivofit; Withings Pulse; Basis Peak |
| Zhang, 2012[68] | Lab | 1 day | 60 | 49 (7) | 62 | Healthy | Activity classification (sedentary, household, walking, and running) | GENE Activ |

COPD: chronic obstructive pulmonary disease; FUP: follow-up time; MVPA: mean to vigorous physical activity; NA: not available.

*Table 3-2: 13 Characteristics of the studies reporting on acceptability.*

| Author, Year | Setting | FUP time | Sample | Mean age (SD) | Female % | Underling disease | Outcome Assessed | Device brand and model |
|---|---|---|---|---|---|---|---|---|
| Boeselt, 2016[69] | Lab and Field | 7 days | 20 | 66 (7) | 15 | COPD | Ease of use and other characteristics | Polar A300 |
| Deka, 2018[70] | Field | 5 days | 46 | 65 (12) | 67 | CHF | Data availability | Fitbit Charge HR |
| Farina, 2019[71] | Field | 2 days | 26 | 80 (6) | 39 | Dementia | Wearing time | GENE Activ |
| | | | 26 | 76 (6) | 73 | Caregivers of patients with dementia | | |
| Fisher, 2016[72] | Field | 7 days | 34 | 69 (min 50, max 86) | NA | Parkinson's disease | Ease of use and other characteristics | AX3 data logger |
| Kaewkannate, 2016[73] | Field | < 1 day | 7 | 31 (0) | 14 | Healthy | Ease of use and other characteristics | Fitbit Flex |
| | | | | | | | | Jawbone Up24 |
| | | | | | | | | Withings Pulse |
| | | | | | | | | Misfit Shine |
| Lahti, 2017[74] | Lab | 120 days | 40 | NA | NA | Schizophrenia | Data availability | Garmin Vivofit |
| Marcoux, 2019[75] | Field | 46 days | 20 | 73 (7) | 20 | Idiopathic Pulmonary Fibrosis | Data availability | Fitbit Flex 2 |
| Naslund, 2015[76] | Field | 80-133 days | 5 | 48 (9) | 90 | Serious mental illness | Wearing time | Nike Fuelband |
| Speier, 2018[77] | Lab | 90 days | 186 | NA | NA | Coronary artery disease | Wearing time | Fitbit Charge HR 2 |
| St-Laurent, 2018[78] | Lab | 1 day | 16 | 33 (4) | 100 | Pregnant | Ease of use and other characteristics | Fitbit Flex |
| Rowlands, 2018[79] | Field | 425 days | 1724 | 13 (1) | 100 | Healthy | Data availability | GENE Activ |

CHF: congestive heart failure; COPD: chronic obstructive pulmonary disease; FUP: follow-up time;

MVPA: mean to vigorous physical activity; NA: not available; SD: standard deviation.

*Table 3-3: Result characteristics of the studies reporting on accuracy.*

| Author, year | Device brand | Device model | Reference standard | Results | Scale of Measure |
|---|---|---|---|---|---|
| Outcome: Active time | | | | | |
| Hernandez-Vicente, 2016[34] | Polar | V800 | ActiTrainer | Mean (SD) bias (Bland-Altman) 32.0 (52.0) min, with mean (SD) 303.95 (93.29) min measured with the reference standard | Activity over 7-days under everyday conditions |
| Outcome: Activity classification (sedentary, household, walking, and running) | | | | | |
| Zhang, 2012[67] | GENE | Activ | Probably direct observation | For different machine learning algorithms (Logistic Regression, Decision Tree, Support Vector Machine, and Bayesian Network) the Incorrect classification rate ranged from 2.71 to 4.44% | 10-12 semi structured activities in lab or outdoor environment while wearing device |
| Outcome: Activity count | | | | | |
| Gironda, 2007[32] | Actiwatch | Score | VICON Motion Analysis System | Pearson's correlation coefficient 0.67-0.88 | Performance on two 15-minute trials of exercise activity prescribed for back-pain rehab. |
| Lawinger, 2015[42] | ActiGraph | GT3X+ | Manual count (video recording) | Correlation r .93, P < .001 "every 4000-vector-magnitude physical activity counts equal 27 arm motions". This 4000 was not pre-specified | Performance on 3 series of tasks: activities of daily living, rehab exercises and passive shoulder range at 5 specified velocities in one lab session. |
| Bruder, 2018[17] | ActivPAL | ActivPAL | 10-camera 3-D Motion analysis system (Vicon-MX3) | Mean difference -40.9 to 30.4 for different activities (95% CI reported) | Performance on two upper limb activities on week apart |
| Outcome: Daily mean activity | | | | | |
| Scott, 2017[54] | GENE | Activ | ActiGraph GT3X+ | Pearson's r 0.88 (95% CI = 0.82–0.93; p = <0.001) | Activity over 7-days under everyday conditions |
| Outcome: Distance | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Gaz, 2018[30] | Fitbit | Charge HR | Measured distance | Mean (SD) difference 0.028 (0.045) to 0.152 (0.124) m | Performance on a free walking or treadmill walking condition. Treadmill walking had pre-determined speeds. |
| Gaz, 2018[30] | Apple | Watch, series not NA | Measured distance | Mean (SD) difference 0.016 (0.05)to 0.037 (0.108) | |
| Gaz, 2018[30] | Garmin | Vivofit 2 | Measured distance | Mean (SD) difference 0.016 (0.028) to 0.107 (0.066) | |
| Gaz, 2018[30] | Jawbone | UP2 | Measured distance | Mean (SD) difference 0.008 (0.049) to 0.086 (0.059) | |
| Huang, 2016[35] | Jawbone | Up24 | Measured distance | Mean (SD) percentage error 5.2 (9.8) during flat ground walking (400 m) | Performance on slow, moderate, and fast walking speeds on treadmill |
| Huang, 2016[35] | Garmin | Vivofit | Measured distance | Mean (SD) percentage error 5.1 (11.4) during flat ground walking (400 m) | Performance on slow, moderate, and fast walking speeds on treadmill |
| Huang, 2016[35] | Fitbit | Flex | Measured distance | Mean (SD) percentage error -12.8 (15.4)% during flat ground walking (400 m) | |
| Wahl, 2017[63] | Beurer | AS80 | Measured distance | MAPE -51.9 to -17.6% | Performance on a treadmill for four 5 minute stages of different velocities, a 5-minute period of intermittent velocity, and a 2.4 km outdoor run and a 2.4 km outdoor run |
| Wahl, 2017[63] | Fitbit | Charge HR | Measured distance | MAPE -29.5 to -13.1% | |
| Wahl, 2017[63] | Fitbit | Charge | Measured distance | MAPE -29.9 to 16% | |
| Wahl, 2017[63] | Garmin | Vivofit | Measured distance | MAPE -25.0 to 23.3% | |
| Wahl, 2017[63] | Garmin | Vivosmart | Measured distance | MAPE -8.1 to 53.5% | |
| Wahl, 2017[63] | Garmin | Vivoactive | Measured distance | MAPE -6.1 to 51.4% | |
| Wahl, 2017[63] | Garmin | Forerunner 920XT | Measured distance | MAPE -3.3 to 26.0% | |
| Wahl, 2017[63] | Xaomi | Mi Band | Measured distance | Not Applicable (too many missing data, not analyzed) | |
| Wahl, 2017[63] | Withings | Pulse | Measured distance | MAPE 0.7 to 58.3% | |

| Outcome: Energy expenditure | | | | | |
|---|---|---|---|---|---|
| Stackpool, 2013[58] | Jawbone | UP | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.20 to 0.87 | First session completed on a treadmill at walking or running speed, selected by the participant. Second session was on elliptical cross-trainer at self-selected speed. Apart of the second session also took place in a gymnasium, where they completed agility ladder drills, basketball throws, and basketball lay-ups. |
| Stackpool, 2013[58] | Nike | Fuelband | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.08 to 0.72 | |
| Stackpool, 2013[58] | Fitbit | Ultra | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.24 to 0.67 | |
| Stackpool, 2013[58] | Adidas | MiCoach | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.55 to 0.81 | |
| Compagnat, 2018[24] | ActiGraph | GT3X+ | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Mean percentage difference 3% for walking subjects, 47% for subjects with wheelchair | Performed four tasks: transfers, manual tasks, walking on flat ground and walking up and down stairs. |
| Hargens, 2017[33] | Fitbit | Charge | ActiGraph GT3x | MAPE 30.6% | Activity over 7-days under everyday conditions |
| Mandigout, 2017[45] | Actical | Actical | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Spearman's r –0.19 (p 0.35) if weared on the plegic side, –0.27 (p 0.23) on the non-plegic side | Performance in various everyday tasks (transfer, walking, etc) within a laboratory setting |
| Mandigout, 2017[45] | ActiGraph | GTX | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Spearman's r 0.08 (p 0.71) if wore on the plegic side, 0.20 (p 0.34) on the non-plegic side | |
| Mandigout, 2017[45] | Fitbit | Charge HR | Indirect calorimetry (Parvo metabolic analyzer) | MAPE (SD) 43.7 (3.4) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, walking various speed and inclines, jogging, and cycling) |
| Price, 2017[49] | Fitbit | One | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias 2.91 (4.35) kcals/min | Walking on a treadmill at varying speeds |
| Price, 2017[49] | Garmin | Vivofit | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias -1.56 (2.34) kcals/min | |
| Price, 2017[49] | Jawbone | UP | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias 18.57 (30.17) kcals/min | |
| Roos, 2017[52] | Suunto | Ambi | Indirect calorimetry | MAPE 21.32 to 41.93% | Aerobic and anaerobic running on a treadmill in a laboratory setting |
| Roos, 2017[52] | Garmin | Forerunner 920XT | Indirect calorimetry | MAPE 11.54 to 49.30% | |
| Roos, 2017[52] | Polar | V800 | Indirect calorimetry | MAPE 10.1 to 39.5% | |

| | | | | | |
|---|---|---|---|---|---|
| Alsubheen, 2016[12] | Garmin | Vivofit | Indirect Calorimetry (Sable Systems International, Las Vegas NV) | Systematically underestimated by 29.5% during treatdmill walking test, p | Performance on treadmill walking tasks, and office activities within a laboratory session, completed in separate sessions on different days |
| Boeselt, 2016[16] | Polar | A300 | BodyMedia SenseWear | Pearson's r 0.74 ($p < 0.01$) | Performance in everyday conditions |
| Choi, 2010[20] | ActiGraph | GT1M | Room calorimeter | Mean (SD) percentage difference: 0.5 (8.0)% | Monitored through a 24-h stay in a laboratory setting. Stay included light activities, eating, sleeping, and participants were encouraged to complete normal day activities during downtime. |
| Chowdhury, 2017[22] | Microsoft | Band | CamNtech Actiheart | MAPE (SD) 34 (10)% | Performance against criterion measurements in both controlled laboratory conditions (simulated activities of daily living and structured exercise) and over a 24-hour period in free-living conditions. |
| Chowdhury, 2017[22] | Apple | Watch, series not NA | CamNtech Actiheart | MAPE (SD) 15 (10)% | |
| Chowdhury, 2017[22] | Jawbone | Up24 | CamNtech Actiheart | MAPE (SD) 30 (11)% | |
| Chowdhury, 2017[22] | Fitbit | Charge | CamNtech Actiheart | MAPE (SD) 16 (8)% | |
| Chowdhury, 2017[22] | Microsoft | Band | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 40 (16)% | |
| Chowdhury, 2017[22] | Apple | Watch, series not NA | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 27 (19)% | |
| Chowdhury, 2017[22] | Jawbone | Up24 | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 36 (14)% | |
| Chowdhury, 2017[22] | Fitbit | Charge | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 36 (22)% | |
| Dondzila, 2018[25] | Fitbit | Charge HR | MET values of treadmill intensities | MAPE -8.4 to 89.2% | Performance on four-five minute stage treadmill tasks in a laboratory session and later in free-living conditions for one day. |
| Dondzila, 2018[25] | Miio | FUSE | MET values of treadmill intensities | MAPE 0 to 44.9% | |

| | | | | | |
|---|---|---|---|---|---|
| Durkalec-Michalski, 2013[27] | ActiGraph | GT1M | Indirect Calorimetery | Overestimated EE at moderate intensity by 60% and underestimated EE by 40% at vigorous intensity. 86% accurate in measuring EE at light intensity in relation to the values measured by indirect calorimetry. | Performance on leisure and exercise activities at various intensities in laboratory and free-living conditions |
| Ferguson, 2015[29] | Misfit | Shine | BodyMedia SenseWear | Mean absolute difference 468, mean (SD) measured with the reference standard = 3005 (569) | Activity under free-living conditions over 48 hours |
| Ferguson, 2015[29] | Jawbone | UP | BodyMedia SenseWear | Mean absolute difference 866, mean (SD) with the reference standard = 3005 (569) | |
| Hernandez-Vicente, 2016[34] | Polar | V800 | Actigraph ActiTrainer | Mean (SD) bias (Bland-Altman) 957.5 (679.9) kcal, with mean (SD) 1,456.48 (731.40) kcals measured with the reference standard | Activity under free-living conditions over 7 days |
| Lemmens, 2018[43] | Phillips | Optical Heart Rate monitor | Indirect calorimetry (portable gas analyser, COSMED K4b2) | Mean percentage error -2.6% | Performance on paced and self-paced exercise activities as well as household activities under laboratory conditions |
| Sirard - Phase 2 (Lab), 2017[56] | Movband | Movband | Indirect calorimetry system (Oxycon Mobile, Carefusion, Inc.) | Spearman's r 0.61 | Performance on structured activities (sitting, self-paced walking, catch, tag, jogging) within a laboratory condition over 2 days |
| Sirard - Phase 2 (Lab), 2017[56] | Sqord | Sqord | Indirect calorimetry system (Oxycon Mobile, Carefusion, Inc.) | Spearman's r 0.87 | |
| Wallen, 2016[64] | Apple | Watch, series not NA | Indirect calorimetry (MetaMax 3B, Cortex, Germany) | Mean (SD) bias (Bland-Altman) -123.1 (55.6) kcal, with index test mean (SD) = 285.7 (50.2) | Completed ~1-hr protocols involving supine and seated rest, walking and running on a treadmill and cycling on an ergometer in a laboratory condition |
| Wallen, 2016[64] | Fitbit | Charge | Indirect calorimetry | Mean (SD) bias (equation reported, since Bland-Altman parameters were systematically biased) 0.61*mean–224.6 (59.1) kcal, with index test mean (SD) = 236.8 (77.0) | |

| | | | | | |
|---|---|---|---|---|---|
| Wallen, 2016[64] | Samsung | Gear S | Indirect calorimetry | Mean (SD) bias (Bland-Altman) -26.1 (24.2) kcal, with index test mean (SD) = 261.4 (47.5) | |
| Wallen, 2016[64] | Miio | Mio alpha | Indirect calorimetry | Mean (SD) bias (equation reported, since Bland-Altman parameters were systematically biased) 0.91*mean -318.77 (84.8) kcal, with index test mean (SD) = 236.8 (77.0) | |
| Woodman, 2017[66] | Garmin | Vivofit | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 44.6 (~8) | Completed 11 activities ranging from sedentary behaviors to vigorous intensities in a laboratory condition over one day |
| Woodman, 2017[66] | Withings | Pulse | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 63.7 (~4.5) | |
| Woodman, 2017[66] | Basis | Peak | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 27.2 (~20) | |
| Imboden, 2018[36] | Fitbit | Flex | Indirect calorimetry | Mean percentage bias = -13% | Participated in an 80-minute protocol of exercises in a laboratory condition |
| Imboden, 2018[36] | Jawbone | Up24 | Indirect calorimetry | Mean percentage bias = -26% | |
| Wahl, 2017[63] | Polar | Loop | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE 5.6 to 56.4% | Performed a running protocol consisting of four 5 min stages of different constant velocities (4.3; 7.2; 10.1; 13.0 km·h−1), a 5 min period of intermittent velocity, and a 2.4 km outdoor run (10.1 km·h−1). |
| Wahl, 2017[63] | Beurer | AS80 | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -48.4 to 17% | |
| Wahl, 2017[63] | Fitbit | Charge HR | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -12.0 to 83.3% | |
| Wahl, 2017[63] | Fitbit | Charge | Indirect calorimetry with portable gas analyzer Metamax | MAPE -4.5 to 75.0% | |

| | | | | | |
|---|---|---|---|---|---|
| | | | 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | | |
| Wahl, 2017[63] | Bodymedia | Sensewear | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -25.3 to -1.4% | |
| Wahl, 2017[63] | Garmin | Vivofit | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -21.3 to 18.7% | |
| Wahl, 2017[63] | Garmin | Vivosmart | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -1.5 to -35.8% | |
| Wahl, 2017[63] | Garmin | Vivoactive | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -4.5 to 36.8% | |
| Wahl, 2017[63] | Garmin | Forerunner 920XT | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -26.6 to -9.2% | |
| Wahl, 2017[63] | Xaomi | Mi Band | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | Not applicable (too many missing data, not analyzed) | |
| Wahl, 2017[63] | Withings | Pulse | Indirect calorimetry with portable gas analyzer Metamax | MAPE -38.9 to -16.9% | |

| | | | | | |
|---|---|---|---|---|---|
| | | | 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | | |
| Dooley, 2017[26] | Apple | Watch, series not NA | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 16.54 (~13) to 210.84 (~96)% | Participants completed a 10-minute seated baseline assessment; separate 4-minute stages of light-, moderate-, and vigorous-intensity treadmill exercises; and a 10-minute seated recovery period in a laboratory setting |
| Dooley, 2017[26] | Fitbit | Charge HR | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 16.85 (~14) to 84.98 (~46)% | |
| Dooley, 2017[26] | Garmin | Forerunner 225 | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 30.77 (~26) to 155.05 (~164)% | |
| Outcome: Heart rate | | | | | |
| Jo, 2016[37] | Basis | Peak K | Standard 12- lead electrocardiograph system (Cosmed C12x; Concord, CA, USA) | Mean(SD) bias (Bland-Altman) -3 (11) bpm | Each participant completed an initial rest period of 15 minutes followed by 5-minute periods of each of the following activities: 60W and 120W cycling, walking, jogging, running, resisted arm raises, resisted lunges, and isometric plank. In between each exercise task was a 5-minute rest period. |
| Jo, 2016[37] | Fitbit | Charge | Standard 12- lead electrocardiograph system (Cosmed C12x; Concord, CA, USA) | Mean(SD) bias (Bland-Altman) -9 (17) bpm | |
| Montoye, 2017[47] | Fitbit | Charge HR | Nonin PureSAT Pulse Oximeter | MAPE (SD) 6.6 (0.6) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, walking various speed and inclines, jogging, and cycling) |
| Dondzila, 2018[25] | Fitbit | Charge HR | Polar heart rate monitor | Trend to report lower mean heart rate values at running speeds of 134.1 m·min-1 and 160.9 m·min-1, compared to the Polar. | Performance on four-five minute stage treadmill tasks in a laboratory session and later in free-living conditions for one day. |
| Dondzila, 2018[25] | Miio | FUSE | Polar heart rate monitor | Mean heart rate values within 1.1 beats·min-1 of the Polar. | |
| Gillinov, 2017[31] | Garmin | Forerunner 235 | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 4.6 (7.7) to 13.7 (16.8)% | Completed exercise protocols on a treadmill, a stationary bicycle, and an elliptical trainer (Tarm movement) in a laboratory setting over one day. |

| | | | | | |
|---|---|---|---|---|---|
| Gillinov, 2017[31] | TomTom | Spark | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 4.5 (5.3) to 6.7 (9.6)% | |
| Gillinov, 2017[31] | Apple | Watch, series not NA | ECG leads, polar H7 chest strap monitor, | MAPE (SD) 3.2 (4.9) to 6.5 (10.8)% | |
| Gillinov, 2017[31] | Fitbit | Blaze | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 5.6 (6.4) to 15.9 (18.2)% | |
| Powierza, 2017[48] | Fitbit | Charge | Electrocardiogram | Mean(SD) bias (Bland-Altman) -6.04 (10.40) bpm | Completed the Buffalo Concussion Treadmill Test in a laboratory setting over one day. |
| Støve, 2019[60] | Garmin | Forerunner | Polar device | Mean difference (SD) 1 (2.3) to 17 (13.36) bpm, with mean (SD) frequence ranging from 59.5 (10.8) to 165.8 (16.4) | Performance during rest and three exercise conditions at submaximal level including cycling, treadmill, walking, running and rapid arm movement in a laboratory setting |
| Thomson, 2019[62] | Apple | Watch, series not NA | ECG | Mean percentage error 2.4 to 5.1% | Measured over performance in different intensity levels of activity, from very light to very rigorous, in a laboratory session over 1 day. |
| Thomson, 2019[62] | Fitbit | Charge HR 2 | Electrocardiogram | Mean percentage error 3.9 to 13.5% | |
| Wallen, 2016[64] | Apple | Watch, series not NA | ECG | Mean (SD) bias (Bland-Altman) -1.3 (4.4) bpm, with index test mean (SD) = 102.0 (14.4) | Completed ~1-hr protocols involving supine and seated rest, walking and running on a treadmill and cycling on an ergometer in a laboratory condition |
| Wallen, 2016[64] | Fitbit | Charge | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -9.3 (8.5) bpm, with index test mean (SD) = 102.0 (14.5) | |
| Wallen, 2016[64] | Samsung | Gear S | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -7.1 (10.3) bpm, with index test mean (SD) = 100.5 (14.6) | |
| Wallen, 2016[64] | Miio | Mio alpha | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -4.3 (7.2) bpm, with index test mean (SD) = 102.0 (14.4) | |

| | | | | | |
|---|---|---|---|---|---|
| Dooley, 2017[26] | Apple | Watch, series not NA | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) 1.4 (~1) to 6.7 (~11)% | Participants completed a 10-minute seated baseline assessment; separate 4-minute stages of light-, moderate-, and vigorous-intensity treadmill exercises; and a 10-minute seated recovery period in a laboratory setting |
| Dooley, 2017[26] | Fitbit | Charge HR | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) 2.4 (~1.5) to 17.0 (~20.0) | |
| Dooley, 2017[26] | Garmin | Forerunner 225 | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) ranging from 7.8 (~17) to 24.38 (~26) | |
| Stiles, 2013[59] | GENE | Activ | Advanced Mechanical Technology Inc. force plate | Sensitivity 97.6%, specificity 75.0%, overall agreement 85.6%, using a cut-off point of 3.125 g | Performed walking (slow, fast, and with bag), floor sweeping, running (slow and fast), jumping (low, G5 cm; high, 95 cm), and box drop (20 cm) in a laboratory session. |
| Stiles, 2013[59] | ActiGraph | GT3X+ | Advanced Mechanical Technology Inc. force plate | Sensitivity 90.5%, specificity 81.3%, overall agreement 85.6 using a pre-specified cut-off cut-off point 2.840 g | |
| Outcome: MVPA | | | | | |
| Semanik, 2020[55] | Fitbit | Flex | ActiGraph GT3X | Mean (SD) difference 18.5 (11.3), with mean (SD) 239.5 (86.2) min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Hargens, 2017[33] | Fitbit | Charge | ActiGraph GT3x | MAPE 46.3% | Activity over 7-days under everyday conditions |
| Scott, 2017[54] | GENE | Activ | ActiGraph GT3X+ | Pearson's r 0.84 (95% CI = 0.77–0.89; p < 0.001) | Activity over 7-days under everyday conditions |
| Boeselt, 2016[16] | Polar | A300 | Bodymedia-SenseWear (SWA) device | Pearson's r -0.25 (p <0.01) | Performance in everyday conditions |
| Ferguson, 2015[29] | Misfit | Shine | Actigraph GT3X+ | Mean absolute difference (MAD) = 15.2, mean (SD) with the reference standard = 58.5 (37.6) | Activity under free-living conditions over 48 hours |
| Ferguson, 2015[29] | Jawbone | UP | Actigraph GT3X+ | Mean absolute difference (MAD) = 18.0, mean (SD) with the reference standard = 58.5 (37.6) | |
| Redenius, 2019[50] | Fitbit | Flex | Actigraph GT3X+ | MAPE (SD) 6.7 (5.7) to 74.3 (12.8)% | Activity over 7-days under everyday conditions |

| | | | | | |
|---|---|---|---|---|---|
| Reid, 2017[51] | Fitbit | Flex | Actigraph GT3X+ | Mean (SD) bias (Bland-Altman) -57.5 (46.4) min/day, with mean 64.6 min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Sirard - Phase 3 (Field), 2017[56] | Movband | Movband | ActiGraph GT3X+ | Spearman's r 0.76 | Activity over 4-days under everyday conditions |
| Sirard - Phase 3 (Field), 2017[56] | Sqord | Sqord | ActiGraph GT3X+ | Spearman's r 0.86 | |
| St-Laurent, 2018[57] | Fitbit | Flex | Actigraph GT3x | Mean (SD) bias (Bland-Altman) 2.4 ± 6.6 (p = 0.21) min/day, with mean (SD) 9.9 (7.5) min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Alharbi, 2016[11] | Fitbit | Flex | Actigraph | Mean percentage error 10% | Activity over 4-days under everyday conditions |
| Imboden, 2018[36] | Fitbit | Flex | ActiGraph GT3X+ | Mean percentage error -65% | Participated in an 80-minute protocol of exercises in a laboratory condition |
| Imboden, 2018[36] | Jawbone | Up24 | ActiGraph GT3X+ | Mean percentage error -35% | |
| Outcome: Physcial activity intensity | | | | | |
| Bulathsinghala, 2014[18] | ActiGraph | GT3X+ | ActiGraph GT3X+ on the waist | Physical activity intensity above the threshold was present in16% of the recorded minutes. Mean Vector Magnitude Unit (VMU - movement in three planes) from the wrist device above the 3000 threshold were 4953 (95% confidence interval (CI), 4850 to 5055), while corresponding VMU from the waist device were 951 (95% CI, 916 to 986). Using a proprietary software equation developed for the waist location, activity intensity above this threshold corresponded to 1.66 metabolic units (METs) (95% CI, 1.55 to 1.77). | Activity over 24 hours under everyday conditions |

| | | | | | |
|---|---|---|---|---|---|
| Outcome: Speed | | | | | |
| Cohen, 2010[23] | ActiGraph | Mini MotionLogger | Actual speed (distance/time) | Mean difference 0.97 mph (95% CI, 0.73 - 2.67) | Completed a standardized sequence of activities that comprised sitting, standing, and walking in laboratory setting |
| Outcome: Step count | | | | | |
| Stackpool, 2013[58] | Jawbone | UP | Manual count | Pearson's r 0.34 to 0.99 | First session completed on a treadmill at walking or running speed, selected by the participant. Second session was on elliptical cross-trainer at self-selected speed. Apart of the second session also took place in a gymnasium, where they completed agility ladder drills, basketball throws, and basketball lay-ups. |
| Stackpool, 2013[58] | Nike | Fuelband | Manual count | Pearson's r 0.17 to 0.98 | |
| Stackpool, 2013[58] | Fitbit | Ultra | Manual count | Pearson's r 0.44 to 0.99 | |
| An, 2017[13] | Fitbit | Flex | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 4.7 to 21.9% in lab, 18.1% on field | Walking/jogging on a treadmill, walking over-ground on an indoor track, and a 24-hour free-living condition |
| An, 2017[13] | Garmin | Vivofit | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 2.4 to 16.5% in lab, 17.8% on field | |
| An, 2017[13] | Polar | Loop | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 9.9 to 23.8% in lab, 26.9% on field | |
| An, 2017[13] | Basis | B1 Band | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 3.1 to 9.0% in lab, 18.4% on field | |
| An, 2017[13] | Misfit | Shine | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 6.3 to 19.3% in lab, 23.3% on field | |

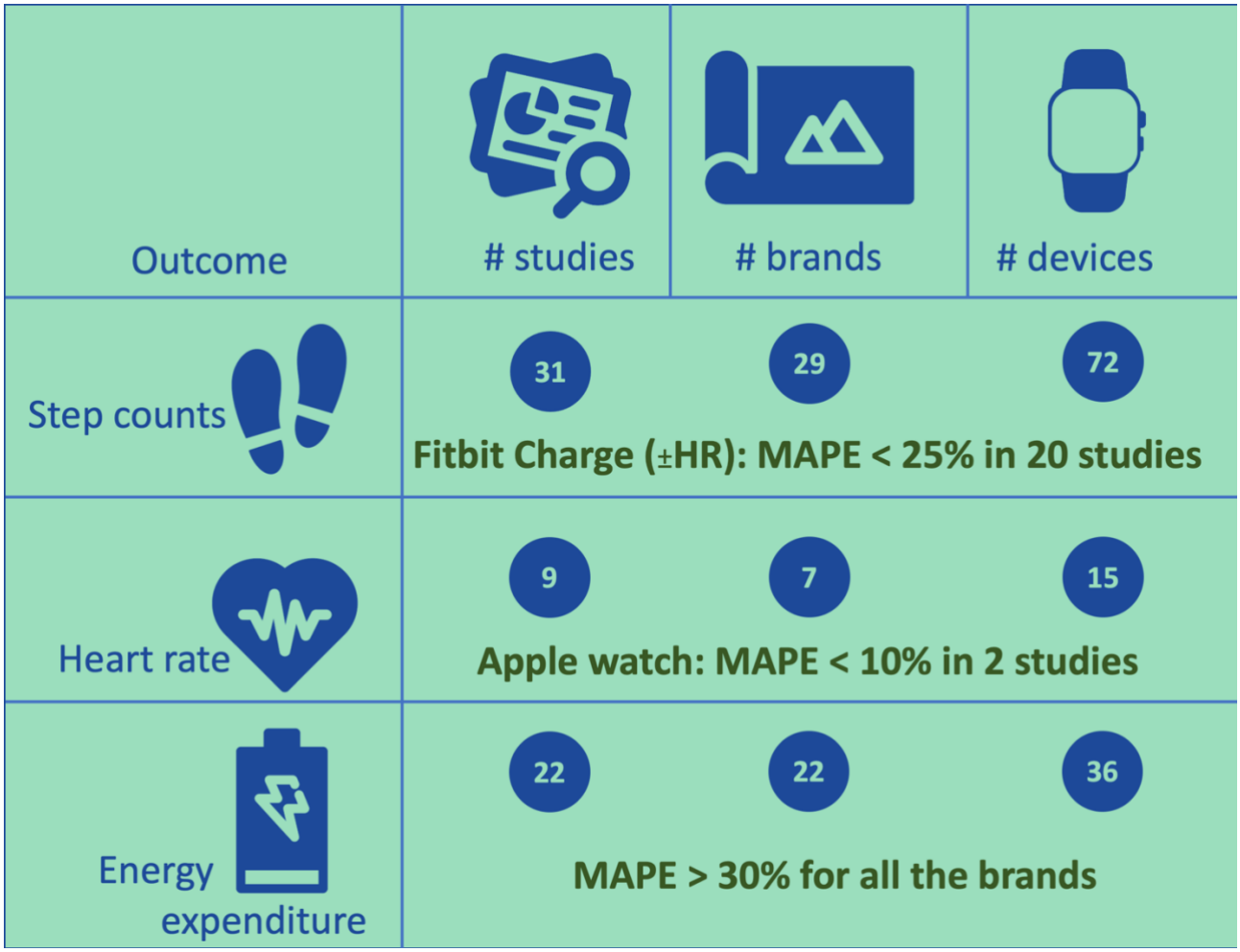
| | | | | | |
|---|---|---|---|---|---|
| An, 2017[13] | Jawbone | UP24 | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 2.9 to 7.0% in lab, 27.9% on field | |
| An, 2017[13] | Nike | FuelBand SE | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 10.2 to 45.0% in lab, 16.0% on field | |
| Gaz, 2018[30] | Fitbit | Charge HR | Manual count (Tally counter) | Mean (SD) difference 21.81 (67.08) to 195.06 (207.94) steps. Max distance 1.6 km. | Performance on a free walking or treadmill walking condition. Treadmill walking had pre-determined speeds |
| Gaz, 2018[30] | Apple | Watch, series not NA | Manual count (Tally counter) | Mean (SD) difference 7.56 (29.61) to 39.44 (151.81) steps. Max distance 1.6 km. | |
| Gaz, 2018[30] | Garmin | Vivofit 2 | Manual count (Tally counter) | Mean (SD) difference 5.09 (8.38) to 98.06 (137.49) steps. Max distance 1.6 km. | |
| Gaz, 2018[30] | Jawbone | UP2 | Manual count (Tally counter) | Mean (SD) difference 16.19 (29.14) to 64 (66.32) steps. Max distance 1.6 km. | |
| Hargens, 2017[33] | Fitbit | Charge | ActiGraph GT3x | MAPE 20.7% | Activity over 7-days under everyday conditions |
| Jones, 2018[38] | Fitbit | Flex | Manual count (video) | MAPE 0-4% | Completed treadmill protocol at jogging and running speeds (8km/h-16km/h) in laboratory settings |
| Lauritzen, 2013[41] | Fitbit | Ultra | Manual count (video) | MAPE (SD) 99.6 (0.8)% | Walking procedure of a straight path over 20m in a laboratory setting |
| Magistro, 2018[44] | ADAMO | Care Watch | Manual count (Tally counter) | MAPE (SD) -17.70 (20.77) % to -1.10 (2.30) % | Performance on randomly ordered tasks: walking slow, normal and fast self-paced speeds, and up/down stairs in a laboratory setting |
| Montoye, 2017[47] | Fitbit | Charge HR | Omron HJ 323u Pedometer (Omron Corp., Osaka, Japan) | MAPE (SD) 9.7% (1.2) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, walking various speed and inclines, jogging, and cycling) |

| | | | | | |
|---|---|---|---|---|---|
| Alsubheen, 2016[12] | Garmin | Vivofit | Kinematics analysis (video camera Sony-HDR-FX1 12X HD, Mini DV Camcorder) | Vivofit systematically underestimated step count only at 0% treadmill inclination | Performance on treadmill walking tasks, and office activities within a laboratory session, completed in separate sessions on different days |
| Falgoust, 2018[28] | Fitbit | Charge HR | Manual count (Tally counter) | Mean difference - 60.8 steps (p 0.01) | Performance on track laps in laboratory settings |
| Falgoust, 2018[28] | Fitbit | Surge | Manual count (Tally counter) | Mean difference -86.0 steps (p 0.004) | |
| Falgoust, 2018[28] | Garmin | Vivoactive HR | Manual count (Tally counter) | Mean difference -19.7 steps (p 0.03) | |
| Blondeel, 2018[15] | Fitbit | Alta | Dynaport Movemonitor (accellerometer) | Mean difference (SD) 773 (829) steps (p=0.009) | Activity over 14-days under everyday conditions |
| Boeselt, 2016[16] | Polar | A300 | BodyMedia SenseWear | Pearson's r 0.96 ($p < 0.01$) | Performance in everyday conditions |
| Burton, 2018[19] | Fitbit | Flex | Manual count (video) by two researchers | Intraclass Correlation (ICC) 0.77 (0.57,0.88) and 0.76 (0.53,0.88) in two 2-minutes walking tests | Two 2-minute walk tests were completed while wearing the fitness trackers. Participants were videoed during each test. Participants were then given one fitness tracker and an accelerometer to wear at home for 14-days. |

**Accuracy**

The accuracy of wrist-wearable activity trackers was assessed in 57 studies on 72 devices from 29 brands. Step count, heart rate, and energy expenditure were the most commonly assessed outcomes in the appraised literature, the results of these outcomes are summarized in in Figure 2-2.

*Figure 3-5: summary of the results for the main accuracy outcomes.*




In this figure, we highlighted the standout device for the most frequently reported outcomes. Icons by Nikhil Bapna, Yoyon Pujiyono, Chintuza, Gregor Cresnar, Andrejs Kirma, Yigit Pinarbasi, from the Noun Project (https://thenounproject.com).[80]

### *Step counts*

31 studies on 72 devices from 29 brands reported data on step counts. The reference standards used were manual count (directly observed or on video, usually with the help of a Tally counter) or automated count through video analysis, an activity tracker (eight different devices), or a photoelectric cell.

The *Actigraph wGT3xBT-BT*, tested against manual count, showed a mean (SD) percentage error: of -41.7 (13.5)%.[22] The *Actigraph GT3x+* showed no statistically significant correlation with the same reference standard.[47]

The *Apple watch* (series not specified) was evaluated in two studies using manual count as the reference standard.[31,65] The mean (SD) difference between the device and the manual count varied from -47 (470) to 39.44 (151.81) steps, in different walking condition.

For the *Fitbit Alta,* the mean (SD) step count was 773 (829) higher (p=0.009) than the one obtained with the reference standard, an accelerometer.[16] For the *Fitbit Charge*, the mean (SD) difference was -59 (704) steps as compared to direct observation.[65] The MAPE for the same device ranged from -4.4% to 20.7%, using different automated step count methods as the reference standard.[34,64,66] The *Fitbit Charge HR* was assessed in 9 studies, using direct observation [20,22,29,31,62] or an automated method of step count as the reference standard [26,41,48,64]. The MAPE ranged from -12.7% and 24.1%. The accuracy of the *Fitbit Flex* in measuring steps was assessed in eleven studies, using manual count[14,20,22,36,37,39,40,47] or an Actigraph device[12,52,58] as the reference standard. The mean percentage error ranged from -23% to 13%. For the *Fitbit One* and *Fitbit Zip*, no statistically significant correlation was found in step counting using direct observation as the reference standard.[47] The correlation coefficient was not reported. For the *Fitbit Surge***,** the mean difference compared to direct observation was -86.0

steps (p 0.004).[29] For the *Fitbit Ultra*, the MAPE (SD) was 99.6% (0.8%)[42] and the Pearson's correlation coefficient against manual count ranged from 0.44 to 0.99 in different exercise conditions.[59]

The accuracy of the *Garmin Vivofit* was assessed in 5 studies,[13,14,36,54,64] with a MAPE ranging from -41% to 18%.[14,54,64]. For the *Vivofit 2***,** one study reported a MAPE of 4%[66] and another one a mean difference (SD) ranging from 5.09 (8.38) to 98.06 (137.49) steps in different walking conditions (over a max distance of 1.6 km).[31]

In a study from Wahl et al., the MAPE against automated step counting using a photoelectric cell as the reference standard, in different exercise type and conditions, ranged from -2.7% to 1.5% for the *Garmin Forerunner 920XT*, from -1.5% to 0.6% for the *Garmin Vivoactiv*, and from -1.1% to -0.3% for the *Garmin Vivosmart*.[64] For the *Garmin Vivoactive HR*, the mean difference against manual step count was -19.7 steps (p 0.03).[29] For the *Garmin Vivosmart HR*, the mean difference (SD) ranged from -39.7 (54.9) to 5.4 (5.8) for different walking speeds and locations (outdoor vs indoor), over a total of 111-686 steps.[41]

For the *Jawbone UP*, the MAPE was -6.73% in one study[66] and the mean absolute difference 806 over an average of 9959 steps in another one.[30] For the *Jawbone UP2*, the mean (SD) difference ranged from 16.19 (29.14) to 64 (66.32) steps for different walking conditions, over a max distance of 1.6 km.[31] For the *Jawbone UP24*, the mean percentage error ranged from -28% to -0.8%.[22,36,37]

For the *Misfit Shine*, the MAPE ranged from -13% to 23%.[14,66]

For the *Miio FUSE*, the MAPE ranged from -5% to -16% at different treadmill speeds,[26] while in another study, the mean percentage error was <5% for the *Miio Mi Band 2*.[62]

For the *Nike Fuelband***,** the mean (SD) percentage error ranged from -34.3 (26.8)% to -16.7 (16.5)%[36] while for the *FuelBand SE* the MAPE ranged from 10.2% to 45.0%.[14]

The MAPE for the *Polar Loop* ranged from -13 to 27% in three studies.[14,64,66] Regarding two other devices from *Polar*, for the *A300*, it was reported a Pearson correlation coefficient of 0.96 ($p < 0.01$),[17] while for the *V800*, the Bland-Altman bias (SD) was equal to 2,487 (2,293) steps/day, over a mean (SD) 10,832 (4,578) steps/day measured with the reference standard.[35]

For the *Withings Pulse*, the MAPE for step count ranged from -16.0% to -0.4%[64] and the accuracy from 97.2-99.9%.[40] All the remaining devices were only used in one study each, and the results are reported in the *supplementary material.*

### *Heart rate*

Nine studies on 15 devices from 7 brands evaluated the accuracy of activity tracking devices to measure the participants' heart rate. The reference standard used were electrocardiography (ECG), pulse oximetry, or another activity tracker (4 different devices).

For the Apple Watch, the MAPE (SD) for measuring heart rate ranged from 1 (~1)% to 7 (~11)%.[27,32]

In the *Fitbit* devices' family, the mean (SD) bias estimated with the Bland-Altman method ranged from -6 (10) to -9 (8) bpm for the *Fitbit Charge*.[38,49,65] For *Fitbit Charge HR* or *HR2*, the MAPE for HR ranged from 2.4 (~1.5)% to 17 (~20)%.[27,48,63] For the *Fitbit Blaze*, the MAPE ranged from 6 (6)% to 16 (18)% for different activities.[32]

### *Active time: time spent in mean to vigorous physical activity and other outcomes*

Thirteen studies on 11 devices from eight brands reported on the time spent being active, most frequently defined as the time spent in mean to vigorous physical activity (MVPA, eleven studies), expressed in min/day. The reference standard for MVPA was another activity tracker (three

different devices). Other outcomes were time spent being active (standing + walking + running), time spent running, or time spent on different types of physical activity, with eachach of these outcomes being reported in only one study.

For the *Fitbit Flex*, the MAPE (SD) for measuring the time spent in MVPA varied from 7 (6)% to 74 (13)%,[51] and the mean percentage error ranged from -65% to 10%.[12,37] All the other devices were only used in one study each, and the results are reported in the *supplementary material.*

### *Intensity of activity: energy expenditure and other outcomes*

Twenty-four studies on 42 devices from 23 brands focused on measuring the intensity of physical activity. The most frequent measure of intensity was energy expenditure (EE), expressed as kcal, evaluated in twenty-two studies. The less frequent measures of intensity included loading rate, and the classification of physical activity (sedentary, household, walking and running). For EE, the reference standard used most commonly was indirect calorimetry (six different instruments). Less common reference standards included EE estimated with other wearable activity trackers (five different devices), estimated based on the treadmill settings, or direct room calorimetry.

Among the *Actigraph* family, the mean percentage difference in the EE compared to the reference standard in people with previous stroke was 3% for walking subjects and 47% for subjects with wheelchair using *Actigraph GT3X+.*[25] The Spearman's correlation coefficient was 0.08 (p 0.71) if worn on the plegic side and 0.20 (p 0.34) on the non-plegic side with *Actigraph GTX.*[46] Using the *ActiGraph GT1M*, the mean (SD) percentage difference was 0.5% (8.0%) in one study,[21] while another one found that the device overestimated EE at moderate intensity by 60% and underestimated EE by 40% at vigorous intensity, while being 86% accurate in measuring EE at light intensity.[28]

For the *Apple Watch***,** the MAPE (SD) for EE ranged from 15% (10%) to 211% (~96%).[23,27]

In the *Fitbit* family, the MAPE from the *Charge* model ranged from -4.5% to 75.0% in different studies[23,34,64] and from -12% to 89% for the *Charge HR*.[26,27,48,64] For the *Fitbit Flex,* a mean percentage bias of -13% was reported.[37] For the *Fitbit One,* one study reported a mean bias of 2.91 (4.35) kcals/min[50], while for *Fitbit Ultra* the Pearson's correlation coefficient ranged from 0.24 to 0.67 for different physical activities.[59]

Among the devices from *Garmin***,** the MAPE for EE ranged from -21% to 45% for the *Vivofit,*[64,67] from -2% to -36% for the *Vivosmart*,[64] and from 5% to 37% for the *Vivoactive*.[64]

For the *Garmin Forerunner*, the MAPE ranged from -27% to 49% for the model *920XT* [53,64] and from 31 (~26) to 155 (~164)% for the model *225.*[27]

In the *Polar* family, the MAPE for EE ranged from 10% to 40% for the *V800* model,[53] with a Bland-Altman bias (SD) of 957.5 (679.9) kcal, when the mean (SD) EE measured with the reference standard was 1,456.48 (731.40) kcals.[35] For the *Polar Loop*, the MAPE for EE ranged from 6% to 56%.[64] The Pearson's correlation coefficient was 0.74 ($p < 0.01$) for the *Polar A300.*[17]

For the *Withings Pulse*, the MAPE for EE ranged from -39% to 64%.[64,67]

### *Less frequently reported outcomes:*

Other outcomes that were evaluated less frequently reported include: distance, reported in three studies on 15 devices from seven brands, always using the measured distance as the reference standard;[31,36,64] speed, reported in one study using one device, with actual speed (on a treadmill) as the reference standard;[24] and activity count, defined as the number of activities (e.g. number of harm movements or of body movements, based on observation or measured acceleration data), reported from four studies on four devices from four different brands, using as the reference standard manual count (video recording), video analysis (automated), or an activity tracker.[18,33,43,55]

### *Risk of bias*

The risk of bias assessment for each outcome is reported in the *supplementary material*. In summary, all the studies were at high/probably high risk of bias for the domain "Patient selection", as they used a convenience sampling technique. Almost all the studies were at low risk of bias for the domains "Index test" and "Reference standard", as the two measurement methods were applied at the same time and interpreted without knowledge of the results obtained with the other method. A small number of studies was identified as high risk for the domain "Flow and timing" based on the high percentage (>25%) of missing data for the index test or reference standard.

**Acceptability**

The acceptability of wrist-wearable activity trackers was assessed in 11 studies on 10 devices from 9 brands.

### *Data availability*

Four studies focused on data availability, expressed as a proportion of time in which the data were available, and a different device was used in each of these studies. The denominator for the proportion could be the study duration or the time spent exercising. Rowlands et al[81] found that data availability was 52% in a pediatric healthy population using *GENE Activ* for 14 months. Deka et al[82] focused on data availability during exercise time. In this study, adult patients with cardiac heart failure activated their *Fitbit Charge HR* in 75% of the exercise sessions (over 5 days), and data was available for 99% of time when activated. Marcoux et al[83] studied the *Fitbit Flex 2* in adults with idiopathic pulmonary fibrosis (for 46 days). Two out of 20 patients did not succeed in activating the device. In the remaining participants, data were available for a mean (SD) of 91%

(20%) of the time. Lahti et al[84] studied the *Garmin Vivofit* in adults with schizophrenia finding available data for 97% of the time (over 4 months).

### *Wearing time*

Three studies reported on the wearing time. Farina et al[85], using *GENE Activ*, found that 89% of the participants with dementia and 86% of their caregivers wore the device for the duration of the study (28 days). Speier et al,[86], using *Fitbit Charge 2*, enrolled participants with coronary artery disease. The median time spent wearing ranged from 44% to 90%, over 90 days. Lastly, for *Nike Fuelband***,** in a study on patients with schizophrenia, the mean (SD) wearing time was 89% (13%) over 80-133 days.[87]

### *Ease of use and other characteristics*

Four studies focused on the ease of use and similar characteristics of wrist-wearing devices. The *Polar A300* was assessed in patients with chronic obstructive pulmonary disease wearing the device for three days, using the Post-Study System Usability Questionnaire (PSSUQ). The PSSUQ calculates a score that ranges from 1 to 7 (the lower the better) for three subdomains.[88] The mean (SD) score was 1.46 (0.23) for system quality, 2.41 (0.53) for information quality, and 3.35 (0.62) for interface quality. The *AX3 data logger* was assessed in persons with Parkinson's disease wearing the device for 7 days.[89] A questionnaire created ad hoc was used for the assessment. Ninety-four percent of participants agreed that they were willing to wear the sensors at home, and 85% agreed that they were willing to wear the sensors in public. However, some participants reported problems with the strap fitting and material (number not reported). The *Fitbit Flex* was assessed with a questionnaire created ad hoc in a study on pregnant women followed for 7 days.[90] The *Fitbit Flex* was reported by 31% to be inconvenient, 6% poorly aesthetic and 12% uncomfortable. Kaewkannate et al[91] asked healthy participants to wear four different devices

over 28 days and compared them using a questionnaire created ad hoc. The *Withings Pulse* had the highest satisfaction score, followed by *Misfit Shine, Jawbone Up24, and Fitbit Flex.*

# Discussion

## Study findings

We systematically reviewed the available evidence on the acceptability and accuracy of wrist-wearable activity tracking devices for measuring physical activity, across different devices and measures. We found substantial heterogeneity between the included studies. The main sources of heterogeneity were the studies' population and setting, the device used, the reference standard, the outcome assessed, and the outcome measures reported.

Acceptability was evaluated in 11 studies on 10 devices from 9 brands. Data availability was $\geq$ 75% for FitBit Charge HR, FitBit Flex 2, and Garmin Vivofit. Data availability is defined as the amount of data captured over a certain time period, which in this case, is over a predetermined duration of each respective study. Data availability can be a measure of how accurate a device is at capturing data when the device is being worn. For example – if an individual wears the device for 8 hours, but only 4 hours of data is available – some questions may be raised on the capability of the device to capture information accurately. The wearing time was 89% for both GENE Activ and Nike Fuelband. Wearing time is defined as the amount of time the device is worn, similarly over a pre-determined duration for each study. For each study, wearing time may have been assessed differently – for example, one study may measure wearing time over a day, whereas another may measure over a week. Both data availability and wearing time can be a deeper look into acceptability, as participants may wear a device more frequently and ultimately, have more data available, if a device is more acceptable. Accuracy was assessed in 57 studies on 72 devices from 29 brands. Among 14 outcomes assessed, step counts, heart rate, and energy expenditure were the ones used more frequently. For step counts, Fitbit Charge (or Charge HR) had a MAPE < 25%

across 20 studies. For heart rate, Apple watch had a MAPE < 10% in 2 studies. For energy expenditure, the MAPE > 30% for all the brands, showing a pour accuracy across devices.

**Comparison with other systematic reviews**

Feehan et al[92] conducted a systematic review on the accuracy of Fitbit devices for measuring physical activity. The review did not specifically focus on wrist wearables, but also included studies using activity trackers worn on other body locations (torso, ankle, or hip). This systematic review reported a good accuracy of FitBit devices in measuring steps, with 46% of the included studies reporting a measurement error within ±3%. Regarding energy expenditure, the authors concluded that "Fitbit devices are unlikely to provide accurate measures of energy expenditure". Studies on heart rate were not included in this review. Evenson et al[93] performed a systematic review focusing on FitBit and Jawbone devices. Similarly, wearing the device on the wrist was not an inclusion criterium. The authors concluded that for step counts the included studies often showed a high correlation (correlation coefficient $\geq 0.80$) between devices from both the brands and the reference standards. The correlation was frequently low for the outcome energy expenditure. Similar to Feehan et al, the outcome heart rate was not included in this systematic review. The results of these systematic reviews are consistent with our findings for the devices and outcomes assessed.

**Strengths and limitations**

The main strengths of our systematic review were the inclusion of all the devices reported in the literature, the reporting on all the outcomes related to acceptability and accuracy, with no restrictions, and the assessment of the risk for bias of the included studies. These characteristics make this review unique for this topic. However, in our systematic review, we decided to exclude studies in which a wearable device was not positioned on a wrist. Some devices can be positioned

both on the wrist or other sites (torso, hip, ankle, arm, or bra) and the acceptability and accuracy can vary for the same device depending on where it is positioned, increasing heterogeneity. [92,93] Therefore, our results cannot be generalized to the acceptability and accuracy of devices worn on sites other than wrists. Regarding the acceptability outcome, screeners were instructed to look for measures of accuracy reported in the title and abstract. We might have missed longitudinal intervention studies reporting on acceptability measures in the full text, but not in the title or abstract. Acceptability is defined and measured in many different way in the literature about wearing devices and about information technology in general, and these definitions are often broad, non-specific.[94] The majority of the included studies used wearing time or data availability as proxies for acceptability. While these metrics have the advantage of being relatively easy to obtain and reproduce, allowing for quantitative comparisons, on the other hand they are only proxies for acceptability, which is a more nuanced concept. For example, one might wonder if wearing time is low because a person only wears the device a few hours each day, or only at weekends, or if they completely stopped wearing it after some time. Moreover, wearing time is more likely to offer valuable information in studies with a long follow up time (FUP), while two out of three studies reporting on this outcome had a FUP of less than one week. Due to the presence of important heterogeneity between studies, we were not able to perform a meta-analysis. Regardless, the comprehensive reporting in this review will allow researchers to assess the available evidence and inform future studies, either to further assess the accuracy of wearable devices or to inform the choice of one device over another to use in interventional studies. To facilitate these choices, we have provided to readers the database with the results of the individual included studies, and we did our best to offer a synthesis of the three outcomes reported most frequently (step counts, heart rate, and energy expenditure).

**Future research**

Further high-quality studies are needed to determine the accuracy and acceptability of wearable devices for measuring physical activity. Given the number of devices available (72 included in this review), it is unlikely that a single study will be able to answer this question. This makes it particularly important to standardize some aspects of these studies, to reduce the heterogeneity between them and allow for meta-syntheses of the results with comparisons across studies, devices, and outcomes. If the heterogeneity was acceptable, a network meta-analysis would also allow to make indirect comparisons. The main sources of heterogeneity that could be controlled are the setting of the study, the population, the reference standard used, and the outcome definition and measure. A first step in this direction would be putting together a task force of experts to issue guidelines on how to report these experiments, similarly to guidelines form the EQUATOR network. A second step would be issuing recommendations on this, starting with accepted reference standards against which devices should be tested for each outcome, the conditions in which the experiment should be conducted, and the way in which the outcomes should be measured and analyzed. Regarding the reference standard, some of these are more accurate than others. Our approach was to take accuracy to mean criterion and convergent validity in this review, but once there is consensus on what are the acceptable reference standard, other comparisons should not be included in a meta synthesis. Regarding the method to report on the accuracy of continuous variables (more common in this field), this is the order of priority that we suggest: MAPE, mean percentage error, mean difference, Bland-Altman mean bias, and measure of correlation as the least preferred. This is because the percentage error gives the reader a better understand of the importance of the error (a mean error of 50 steps is much more relevant if the total step count was 100 than if it was 10,000). We preferred the MAPE over the

mean absolute error because not using the absolute value there is a risk of negative and positive errors balancing each other, with the risk of overestimating the accuracy. We prefer the mean difference over the Bland-Altman mean bias because in an accuracy study the reference standard is supposed to be more accurate than the index text, and therefore the latter should be tested against the former, not against their mean. In the case of the acceptability outcome, consensus should be reached also about how to define and measure it. For example, defining a minimum set of outcomes to be reported might help in this context. This might include reporting the percentage of abandonment over time. Furthermore, as new devices become available, their acceptability and accuracy should also be tested, as they could differ from the acceptability and accuracy of other devices, even ones produced by the same company. Regarding the choice of the device to use in interventional studies, for example in studies that aim at increasing physical activity in a certain population, there is no one-device-fits-all answer. This choice should be based on the available data on acceptability and accuracy and be tailored to the outcome to measure. In a study with step count as the main outcome, Fitbit Charge and Charge HR might be appropriate choices. Apple watch might be preferred if the main outcome is heart rate. Active time was most often measured through time spent in MVPA, and Fitbit Flex is the only device that was used in three studies, showing good results in two of these. Regarding EE, we don't fill comfortable suggesting the use of any device over another one, based on the current evidence, as the accuracy was poor across devices. Probably the decision should be driven by the other outcomes used. Broader recommendations should be issued in guidelines from a panel of expert using this systematic review as a knowledge base.

## Conclusions

We reported on the acceptability and accuracy of 72 wrist-wearable devices for measuring physical activity produced by 29 companies. Fitbit Charge and Charge HR were consistently shown to have a good accuracy for step counts and Apple watch for measuring heart rate. None of the tested devices proved to be accurate in measuring energy expenditure. Efforts should be made to reduce the heterogeneity between studies.

## Declarations

### Funding

We received no funding for this systematic review.

### Conflicts of interest/Competing interests

The authors have no conflict of interest to declare.

### Ethics approval

The systematic review was based on published data and therefore did not require a submission to a Research Ethics Board (REB).

### Availability of data and material

The majority of the data that support the findings of this study are available in the appendix. The full dataset can be made available, upon reasonable request.

### Code availability

Not applicable to this systematic review, as no quantitative data synthesis was performed.

### Authors' contributions

FG and AI conceived this review. FG is the guarantor of the review and drafted the manuscript. TN and VBD developed the search strategy. FG, NN, VBD, APB, and DP screened the articles and extracted the data. All the authors read, provided feedback and approved the final manuscript.

### Abbreviations

Applications: apps

mHealth: mobile health

PRISMA extension for Scoping Reviews: PRISMA-ScR

MAPE: mean absolute percentage

## References

1 Seifert A, Schlomann A, Rietz C, *et al.* The use of mobile devices for physical activity tracking in older adults' everyday life. *DIGITAL HEALTH* 2017;**3**:205520761774008. doi:10.1177/2055207617740088

2 Ainsworth BE. How do i measure physical activity in my patients? Questionnaires and objective methods. Br J Sports Med. 2009;**43**:6–9. doi:10.1136/bjsm.2008.052449

3 Rütten A, Ziemainz H, Schena F, *et al.* Using different physical activity measurements in eight European countries. Results of the European Physical Activity Surveillance System (EUPASS) time series survey. *Public Health Nutr* 2003;**6**:371–6. doi:10.1079/phn2002450

4 Rothwell PM. Analysis of agreement between measurements of continuous variables: General principles and lessons from studies of imaging of carotid stenosis. Journal of Neurology. 2000;**247**:825–34. doi:10.1007/s004150070068

5 P R, CS P, R A. Common pitfalls in statistical analysis: Measures of agreement. *Perspectives in clinical research* 2017;**8**:187–91. doi:10.4103/PICR.PICR_123_17

6 Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Family medicine* 2005;**37**:360–3.

7 Rothwell PM. Analysis of agreement between measurements of continuous variables: General principles and lessons from studies of imaging of carotid stenosis. Journal of Neurology. 2000;**247**:825–34. doi:10.1007/s004150070068

8 Dillon A, Morris MG. User acceptance of new information technology: theories and models. In: *Annual Review of Information Science and Technology*. Medford, N.J.: Information Today 1996.

9 Shin G, Jarrahi MH, Fei Y, *et al.* Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. Journal of Biomedical Informatics. 2019;**93**. doi:10.1016/j.jbi.2019.103153

10 Ouzzani M, Hammady H, Fedorowicz Z, *et al.* Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016;**5**:210. doi:10.1186/s13643-016-0384-4

11 Whiting PF, Rutjes AWS, Westwood ME, *et al.* Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;**155**:529–36. doi:10.7326/0003-4819-155-8-201110180-00009

12 Alharbi M, Bauman A, Neubeck L, *et al.* Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population. *European Journal of Preventive Cardiology* 2016;**23**:1476–85. doi:10.1177/2047487316634883

13 Alsubheen SA, George AM, Baker A, *et al.* Accuracy of the vivofit activity tracker. *Journal of Medical Engineering and Technology* 2016;**40**:298–306. doi:10.1080/03091902.2016.1193238

14 An HS, Jones GC, Kang SK, *et al.* How valid are wearable physical activity trackers for measuring steps? *European Journal of Sport Science* 2017;**17**:360–8. doi:10.1080/17461391.2016.1255261

15 An HS, Kim Y, Lee JM. Accuracy of inclinometer functions of the activPAL and ActiGraph GT3X+: A focus on physical activity. *Gait and Posture* 2017;**51**:174–80. doi:10.1016/j.gaitpost.2016.10.014

16 Blondeel A, Loeckx M, Rodrigues F, *et al.* Wearables to Coach Physical Activity in Patients with COPD: Validity and Patient's Experience. In: *D51. PHYSIOLOGY AND PHYSICAL ACTIVITY IN PULMONARY REHABILITATION*. American Thoracic Society 2018. A7072--A7072.

17 Boeselt T, Spielmanns M, Nell C, *et al.* Validity and usability of physical activity monitoring in patients with Chronic Obstructive Pulmonary Disease (COPD). *PLoS ONE* 2016;**11**. doi:10.1371/journal.pone.0157229

18 Bruder AM, McClelland JA, Shields N, *et al.* Validity and reliability of an activity monitor to quantify arm movements and activity in adults following distal radius fracture. *Disability and Rehabilitation* 2018;**40**:1318–25. doi:10.1080/09638288.2017.1288764

19 Bulathsinghala C, Tejada J, Zu Wallack R. Validating Output From an Activity Monitor Worn on the Wrist in Patients With COPD. *Chest* 2014;**146**:25A.

20 Burton E, Hill KD, Lautenschlager NT, *et al.* Reliability and validity of two fitness tracker devices in the laboratory and home environment for older community-dwelling people. *BMC Geriatrics* 2018;**18**. doi:10.1186/s12877-018-0793-4

21 Choi L, Chen KY, Acra SA, *et al.* Distributed lag and spline modeling for predicting energy expenditure from accelerometry in youth. *Journal of Applied Physiology* 2010;**108**:314–27. doi:10.1152/japplphysiol.00374.2009

22 Chow JJ, Thom JM, Wewege MA, *et al.* Accuracy of step count measured by physical activity monitors: The effect of gait speed and anatomical placement site. *Gait and Posture* 2017;**57**:199–203. doi:10.1016/j.gaitpost.2017.06.012

23 Chowdhury EA, Western MJ, Nightingale TE, *et al.* Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors. *PLoS ONE* 2017;**12**. doi:10.1371/journal.pone.0171720

24 Cohen MD, Cutaia M. A novel approach to measuring activity in chronic obstructive pulmonary disease: Using 2 activity monitors to classify daily activity. *Journal of Cardiopulmonary Rehabilitation and Prevention* 2010;**30**:186–94. doi:10.1097/HCR.0b013e3181d0c191

25 Compagnat M, Mandigout S, Chaparro D, *et al.* Validity of the Actigraph GT3x and influence of the sensor positioning for the assessment of active energy expenditure during four activities of daily living in stroke subjects. *Clinical Rehabilitation* 2018;**32**:1696–704. doi:10.1177/0269215518788116

26 Dondzila CJ, Lewis C, Lopez JR, *et al.* Congruent accuracy of wrist-worn activity trackers during controlled and free-living conditions. *International Journal of Exercise Science* 2018;**11**:575–84.

27 Dooley EE, Golaszewski NM, Bartholomew JB. Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices. *JMIR mHealth and uHealth* 2017;**5**. doi:10.2196/mhealth.7043

28 Durkalec-Michalski K, Woźniewicz M, Bajerska J, *et al.* Comparison of accuracy of various non-calorimetric methods measuring energy expenditure at different intensities. *Human Movement* 2013;**14**:161–7.

29 Falgoust B, Handwerger B, Rouly L, *et al.* Accuracy of wearable activity monitors in the measurement of step count and distance. *Cardiopulmonary Physical Therapy Journal* 2018;**29**:36–36.

30 Ferguson T, Rowlands A V., Olds T, *et al.* The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study.

*International Journal of Behavioral Nutrition and Physical Activity* 2015;**12**. doi:10.1186/s12966-015-0201-9

31 Gaz D V., Rieck TM, Peterson NW, *et al.* Determining the Validity and Accuracy of Multiple Activity-Tracking Devices in Controlled and Free-Walking Conditions. *American Journal of Health Promotion* 2018;**32**:1671–8. doi:10.1177/0890117118763273

32 Gillinov AM, Etiwy M, Gillinov S, *et al.* Variable accuracy of commercially available wearable heart rate monitors. *Journal of the American College of Cardiology* 2017;**69**:336.

33 Gironda RJ, Lloyd J, Clark ME, *et al.* Preliminary evaluation of reliability and criterion validity of Actiwatch-Score. *Journal of Rehabilitation Research and Development* 2007;**44**:223–30. doi:10.1682/JRRD.2006.06.0058

34 Hargens TA, Deyarmin KN, Snyder KM, *et al.* Comparison of wrist-worn and hip-worn activity monitors under free living conditions. *Journal of Medical Engineering and Technology* 2017;**41**:200–7. doi:10.1080/03091902.2016.1271046

35 Hernández-Vicente HVA, Santos-Lozano AL, De Cocker K, *et al.* Validation study of Polar V800 accelerometer. *Annals of Translational Medicine* 2016;**4**. doi:10.21037/atm.2016.07.16

36 Huang Y, Xu J, Yu B, *et al.* Validity of FitBit, Jawbone UP, Nike+ and other wearable devices for level and stair walking. *Gait and Posture* 2016;**48**:36–41. doi:10.1016/j.gaitpost.2016.04.025

37 Imboden MT, Nelson MB, Kaminsky LA, *et al.* Comparison of four Fitbit and Jawbone activity monitors with a research-grade ActiGraph accelerometer for estimating physical activity and energy expenditure. *British journal of sports medicine* 2018;**52**:844–50. doi:10.1136/bjsports-2016-096990

38 Jo, Edward and Lewis, Kiana and Directo, Dean and Kim, Michael J and Dolezal BA. Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking - PubMed. Journal of sports science & medicine. 2016;:540.https://pubmed.ncbi.nlm.nih.gov/27803634/ (accessed 11 Dec 2020).

39 Jones D, Crossley K, Dascombe B, *et al.* VALIDITY AND RELIABILITY OF THE FITBIT FLEX™ AND ACTIGRAPH GT3X+ AT JOGGING AND RUNNING SPEEDS. *International journal of sports physical therapy* 2018;**13**:860–70.

40 Kaewkannate K, Kim S. A comparison of wearable fitness devices. *BMC Public Health* 2016;**16**. doi:10.1186/s12889-016-3059-0

41 Lamont RM, Daniel HL, Payne CL, *et al.* Accuracy of wearable physical activity trackers in people with Parkinson's disease. *Gait & posture* 2018;**63**:104–8.

42 Lauritzen J, Muñoz A, Sevillano Ramos JL, *et al.* The usefulness of activity trackers in elderly with reduced mobility: a case study. *Studies in Health Technology and Informatics Volume 192: MEDINFO 2013* 2013.

43 Lawinger E, Uhl TL, Abel M, *et al.* Assessment of accelerometers for measuring upper-extremity physical activity. *Journal of sport rehabilitation* 2015;**24**:236–43.

44 Lemmens PMC, Sartor F, Cox LGE, *et al.* Evaluation of an activity monitor for use in pregnancy to help reduce excessive gestational weight gain. *BMC Pregnancy and Childbirth* 2018;**18**. doi:10.1186/s12884-018-1941-8

45 Magistro D, Brustio PR, Ivaldi M, *et al.* Validation of the ADAMO Care Watch for step counting in older adults. *PLoS ONE* 2018;**13**. doi:10.1371/journal.pone.0190753

46 Mandigout S, Lacroix J, Ferry B, *et al.* Can energy expenditure be accurately assessed using accelerometry-based wearable motion detectors for physical activity monitoring in

post-stroke patients in the subacute phase? *European Journal of Preventive Cardiology* 2017;**24**:2009–16. doi:10.1177/2047487317738593

47 Manning MR, Tune BN, Völgyi E, *et al.* Agreement of activity monitors during treadmill walking in teenagers with severe obesity. *Pediatric Exercise Science* 2016;**28**.

48 Montoye AHK, Mitrzyk JR, Molesky MJ. Comparative accuracy of a wrist-worn activity tracker and a smart shirt for physical activity assessment. *Measurement in Physical Education and Exercise Science* 2017;**21**:201–11.

49 Powierza CS, Clark MD, Hughes JM, *et al.* Validation of a Self-Monitoring Tool for Use in Exercise Therapy. *PM and R* 2017;**9**:1077–84. doi:10.1016/j.pmrj.2017.03.012

50 Price K, Bird SR, Lythgo N, *et al.* Validation of the Fitbit One, Garmin Vivofit and Jawbone UP activity tracker in estimation of energy expenditure during treadmill walking and running. *Journal of Medical Engineering and Technology* 2017;**41**:208–15. doi:10.1080/03091902.2016.1253795

51 Redenius N, Kim Y, Byun W. Concurrent validity of the Fitbit for assessing sedentary behavior and moderate-to-vigorous physical activity. *BMC Medical Research Methodology* 2019;**19**. doi:10.1186/s12874-019-0668-1

52 Reid RER, Insogna JA, Carver TE, *et al.* Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. *Journal of Science and Medicine in Sport* 2017;**20**:578–82. doi:10.1016/j.jsams.2016.10.015

53 Roos L, Taube W, Beeler N, *et al.* Validity of sports watches when estimating energy expenditure during running. *BMC Sports Science, Medicine and Rehabilitation* 2017;**9**. doi:10.1186/s13102-017-0089-6

54 Schaffer SD, Holzapfel SD, Fulk G, *et al.* Step count accuracy and reliability of two activity tracking devices in people after stroke. *Physiotherapy Theory and Practice* 2017;**33**:788–96. doi:10.1080/09593985.2017.1354412

55 Scott JJ, Rowlands A V., Cliff DP, *et al.* Comparability and feasibility of wrist- and hip-worn accelerometers in free-living adolescents. *Journal of Science and Medicine in Sport* 2017;**20**:1101–6. doi:10.1016/j.jsams.2017.04.017

56 Semanik P, Lee J, Pellegrini CA, *et al.* Comparison of Physical Activity Measures Derived From the Fitbit Flex and the ActiGraph GT3X+ in an Employee Population With Chronic Knee Symptoms. *ACR Open Rheumatology* 2020;**2**:48–52. doi:10.1002/acr2.11099

57 Sirard JR, Masteller B, Freedson PS, *et al.* Youth oriented activity trackers: Comprehensive laboratoryand field-based validation. *Journal of Medical Internet Research* 2017;**19**. doi:10.2196/jmir.6360

58 St-Laurent A, Mony MM, Mathieu M, *et al.* Validation of the Fitbit Zip and Fitbit Flex with pregnant women in free-living conditions. *Journal of Medical Engineering and Technology* 2018;**42**:259–64. doi:10.1080/03091902.2018.1472822

59 Stackpool CM. Accuracy of various activity trackers in estimating steps taken and energy expenditure. 2013.

60 Stiles VH, Griew PJ, Rowlands A V. Use of accelerometry to classify activity beneficial to bone in premenopausal women. *Medicine and Science in Sports and Exercise* 2013;**45**:2353–61. doi:10.1249/MSS.0b013e31829ba765

61 Støve MP, Haucke E, Nymann ML, *et al.* Accuracy of the wearable activity tracker Garmin Forerunner 235 for the assessment of heart rate during rest and activity. *Journal of Sports Sciences* 2019;**37**:895–901. doi:10.1080/02640414.2018.1535563

62 Tam KM, Cheung SY. Validation of electronic activity monitor devices during treadmill walking. *Telemedicine and e-Health* 2018;**24**:782–9. doi:10.1089/tmj.2017.0263

63 Thomson EA, Nuss K, Comstock A, *et al.* Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *Journal of Sports Sciences* 2019;**37**:1411–9. doi:10.1080/02640414.2018.1560644

64 Wahl Y, Düking P, Droszez A, *et al.* Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Frontiers in Physiology* 2017;**8**. doi:10.3389/fphys.2017.00725

65 Wallen MP, Gomersall SR, Keating SE, *et al.* Accuracy of heart rate watches: Implications for weight management. *PLoS ONE* 2016;**11**. doi:10.1371/journal.pone.0154420

66 Wang L, Liu T, Wang Y, *et al.* Evaluation on Step Counting Performance of Wristband Activity Monitors in Daily Living Environment. *IEEE Access* 2017;**5**:13020–7. doi:10.1109/ACCESS.2017.2721098

67 Woodman JA, Crouter SE, Bassett DR, *et al.* Accuracy of Consumer Monitors for Estimating Energy Expenditure and Activity Type. *Medicine and Science in Sports and Exercise* 2017;**49**:371–7. doi:10.1249/MSS.0000000000001090

68 Zhang S, Murray P, Zillmer R, *et al.* Activity classification using the genea: Optimum sampling frequency and number of axes. *Medicine and Science in Sports and Exercise* 2012;**44**:2228–34. doi:10.1249/MSS.0b013e31825e19fd

69 Boeselt T, Spielmanns M, Nell C, *et al.* Validity and usability of physical activity monitoring in patients with Chronic Obstructive Pulmonary Disease (COPD). *PLoS ONE* 2016;**11**. doi:10.1371/journal.pone.0157229

70 Deka P, Pozehl B, Norman JF, *et al.* Feasibility of using the Fitbit ® Charge HR in validating self-reported exercise diaries in a community setting in patients with heart failure. *European Journal of Cardiovascular Nursing* 2018;**17**:605–11. doi:10.1177/1474515118766037

71 Farina N, Sherlock G, Thomas S, *et al.* Acceptability and feasibility of wearing activity monitors in community-dwelling older adults with dementia. *International Journal of Geriatric Psychiatry* 2019;**34**:617–24. doi:10.1002/gps.5064

72 Fisher JM, Hammerla NY, Rochester L, *et al.* Body-Worn Sensors in Parkinson's Disease: Evaluating Their Acceptability to Patients. *Telemedicine and e-Health* 2016;**22**:63–9. doi:10.1089/tmj.2015.0026

73 Kaewkannate K, Kim S. A comparison of wearable fitness devices. *BMC Public Health* 2016;**16**. doi:10.1186/s12889-016-3059-0

74 Lahti A, White D, Katiyar N, *et al.* 325. Clinical Utility Study Towards the Use of Continuous Wearable Sensors and Patient Reported Surveys for Relapse Prediction in Patients at High Risk of Relapse in Schizophrenia. *Biological Psychiatry* 2017;**81**:S133.

75 Marcoux V, Fell CD, Johannson KAM. Daily Activity Trackers and Home Spirometry in Patients with Idiopathic Pulmonary Fibrosis. In: *A41. ILD SCIENTIFIC ABSTRACTS: DIAGNOSIS, OUTCOMES, AND PPFE*. American Thoracic Society 2018. A1586--A1586.

76 Naslund JA, Aschbrenner KA, Barre LK, *et al.* Feasibility of popular m-health technologies for activity tracking among individuals with serious mental Illness. *Telemedicine and e-Health* 2015;**21**:213–6. doi:10.1089/tmj.2014.0105

77 Speier W, Dzubur E, Zide M, *et al.* Evaluating utility and compliance in a patient-based eHealth study using continuous-Time heart rate and activity trackers. *Journal of the American Medical Informatics Association* 2018;**25**:1386–91. doi:10.1093/jamia/ocy067

78 St-Laurent A, Mony MM, Mathieu M, *et al.* Validation of the Fitbit Zip and Fitbit Flex with pregnant women in free-living conditions. *Journal of Medical Engineering and Technology* 2018;**42**:259–64. doi:10.1080/03091902.2018.1472822

79 Rowlands A V., Harrington DM, Bodicoat DH, *et al.* Compliance of Adolescent Girls to Repeated Deployments of Wrist-Worn Accelerometers. *Medicine and Science in Sports and Exercise* 2018;**50**:1508–17. doi:10.1249/MSS.0000000000001588

80 Noun Project: Free Icons & Stock Photos for Everything. https://thenounproject.com/ (accessed 11 Sep 2021).

81 Rowlands A V., Harrington DM, Bodicoat DH, *et al.* Compliance of Adolescent Girls to Repeated Deployments of Wrist-Worn Accelerometers. *Medicine and Science in Sports and Exercise* 2018;**50**:1508–17. doi:10.1249/MSS.0000000000001588

82 Deka P, Pozehl B, Norman JF, *et al.* Feasibility of using the Fitbit ® Charge HR in validating self-reported exercise diaries in a community setting in patients with heart failure. *European Journal of Cardiovascular Nursing* 2018;**17**:605–11. doi:10.1177/1474515118766037

83 Marcoux V, Fell CD, Johannson KAM. Daily Activity Trackers and Home Spirometry in Patients with Idiopathic Pulmonary Fibrosis. In: *A41. ILD SCIENTIFIC ABSTRACTS: DIAGNOSIS, OUTCOMES, AND PPFE*. American Thoracic Society 2018. A1586--A1586.

84 Lahti A, White D, Katiyar N, *et al.* 325. Clinical Utility Study Towards the Use of Continuous Wearable Sensors and Patient Reported Surveys for Relapse Prediction in

Patients at High Risk of Relapse in Schizophrenia. *Biological Psychiatry* 2017;**81**:S133. doi:10.1016/j.biopsych.2017.02.340

85 Farina N, Sherlock G, Thomas S, *et al.* Acceptability and feasibility of wearing activity monitors in community-dwelling older adults with dementia. *International Journal of Geriatric Psychiatry* 2019;**34**:617–24. doi:10.1002/gps.5064

86 Speier W, Dzubur E, Zide M, *et al.* Evaluating utility and compliance in a patient-based eHealth study using continuous-Time heart rate and activity trackers. *Journal of the American Medical Informatics Association* 2018;**25**:1386–91. doi:10.1093/jamia/ocy067

87 Naslund JA, Aschbrenner KA, Barre LK, *et al.* Feasibility of popular m-health technologies for activity tracking among individuals with serious mental Illness. *Telemedicine and e-Health* 2015;**21**:213–6. doi:10.1089/tmj.2014.0105

88 Boeselt T, Spielmanns M, Nell C, *et al.* Validity and usability of physical activity monitoring in patients with Chronic Obstructive Pulmonary Disease (COPD). *PLoS ONE* 2016;**11**. doi:10.1371/journal.pone.0157229

89 Fisher JM, Hammerla NY, Rochester L, *et al.* Body-Worn Sensors in Parkinson's Disease: Evaluating Their Acceptability to Patients. *Telemedicine and e-Health* 2016;**22**:63–9. doi:10.1089/tmj.2015.0026

90 St-Laurent A, Mony MM, Mathieu M, *et al.* Validation of the Fitbit Zip and Fitbit Flex with pregnant women in free-living conditions. *Journal of Medical Engineering and Technology* 2018;**42**:259–64. doi:10.1080/03091902.2018.1472822

91 Kaewkannate K, Kim S. A comparison of wearable fitness devices. *BMC Public Health* 2016;**16**. doi:10.1186/s12889-016-3059-0

92 Feehan LM, Geldman J, Sayre EC, *et al.* Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. JMIR mHealth and uHealth. 2018;**6**. doi:10.2196/10527

93 Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. International Journal of Behavioral Nutrition and Physical Activity. 2015;**12**. doi:10.1186/s12966-015-0314-1

94 Nadal C, Doherty G, Sas C. Technology acceptability, acceptance and adoption-definitions and measurement. In: *2019 CHI Conference on Human Factors in Computing Systems*. 2019.

# Chapter 4 – validating a possible predictive model

## WAPPS next? Validation of a prediction model for the risk of bleeding in people with Hemophilia, using PK data from the Web-Accessible Population Pharmacokinetic Service (WAPPS-Hemo) and the Canadian Bleeding Disorders Registry (CBDR).

This article has not yet been submitted for publication, pending the publication of a parent article (reported in the supplement). Therefore, the content of this chapter is currently embargoed.

### Authors

Federico Germini[1, 2], Pierre Chelle[3], Kerstin de Wit,[1,4] Giorgio Gosti[5], Dagmar Hajducek[3], Emma Iserman[1], Arun Keepanasseril,[1] Davide Matino,[2] Sameer Parpia,[1, 6] Lehana Thabane,[1,7-10] Andrea Edginton[1], Alfonso Iorio[3].

### Affiliations

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada;

[2] Department of Medicine, McMaster University, Hamilton, ON, Canada;

[3] School of Pharmacy, University of Waterloo, Waterloo, Ontario, Canada;

[4] Department of Emergency Medicine, Queen's University, Kingston, ON, Canada;

[5] Center for Life NanoScience, Istituto Italiano di Tecnologia, Rome, Italy;

[6] Department of Oncology, McMaster University, Hamilton, ON, Canada

[7] Population Health Research Institute, Hamilton Health Sciences, Hamilton, ON, Canada;

[8] Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, Hamilton, ON, Canada;

[9] Departments of Paediatrics and Anaesthesia, McMaster University, Hamilton, ON, Canada;

[10] Centre for Evaluation of Medicine, St Joseph's Healthcare, Hamilton, ON, Canada.

## Corresponding author

Dr. Federico Germini, Health Information Research Unit (HIRU), Communication Research Laboratory (CRL), McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada: germinif@mcmaster.ca

## Background

In the introduction to this thesis (Chapter 1), we outlined how a risk assessment model (RAM) to predict the risk of bleeding in people leaving with hemophilia (PWH) would be helpful to patients, physicians, and policymakers. In our systematic review of the literature on RAMs with this scope (Chapter 2), we did not find any and decided to perform a systematic review of risk factors for bleeding in PWH on regular prophylaxis.[1] We found that plasma factor levels, bleeding history, and physical activity should be considered in the derivation analysis when building a RAM for bleeding in PWH, and the role of other risk factors, including antithrombotic treatment and obesity, should be explored.[1] Having identified this gap in the literature, the next natural step would be deriving and validating such a RAM. Unfortunately, in the Canadian Registry of Bleeding Disorders (CBDR), we do not systematically collect data on physical activity. Our systematic review on the acceptability and accuracy of wrist-wearable activity tracking devices (Chapter 3) had the aim of identifying one or more devices that are best suited for measuring physical activity in PWH, to later collect and use these data to build a RAM. However, in parallel to the work described above, some of us, led by Dr. Pierre Chelle, were collaborating with a group based in the United States of America, using data from CBDR, WAPPS, and the American Thrombosis and Hemostasis Network (ATHN), to further explore the association between factor levels and bleeding episodes. A manuscript describing the results of this project, still unpublished, is reported in the supplementary material. The main output of the project was a repeated time-to-event (RTTE) model that allowed estimating the bleeding risk based on the factor levels over time, derived using individual pop-PK data profiles and infusion logs. This model could be used as a RAM to predict the individual risk of bleeding in PWH. If this

RAM proved to be accurate, there would be no need to reinvent the wheel in producing a new one.

**Objectives:**

The primary objective of this study was to validate the performance of a recently derived RAM for the prediction of the risk of bleeding in PWH A with an available individual pop-PK profile (i.e., the same type of participants from which the model was derived). The secondary objective was the validation of the same RAM in different populations, namely PWH A without an individual pop-PK profile, and PWH B, with and without an available individual pop-PK profile.

## Methods

This study was a pre-specified analysis of prospectively routinely collected data, reported following the TRIPOD[2] and RECORD[3] statements.

**Research question:**

What is the performance of a previously derived RAM for the prediction of the risk for bleeding in PWH A and B, with and without an available individual pop-PK profile?

**Source of data**

Data have been collected using WAPPS-Hemo and CBDR. These databases are already linked in the back end. Therefore, no linking procedure was needed.

**Participants**

All Canadian PWH A or B, on regular prophylaxis with a factor concentrate, actively recording data in CBDR. The study population was classified into four categories:

1. PWH A with an individual pop-PK profile available on WAPPS-Hemo.
2. PWH A without an individual pop-PK profile available on WAPPS-Hemo.
3. PWH B with an individual pop-PK profile available on WAPPS-Hemo.
4. PWH B without an individual pop-PK profile available on WAPPS-Hemo.

For the first category, we used data from January 1$^{st}$, 2021, to December 31$^{st}$, 2021, as older data have been used for the model derivation. For categories #2-4, we used data from January 1$^{st}$, 2018 to December 31$^{st}$, 2021, as none of these data have been used for the model derivation. Only participants actively recording their treatments and bleeds were included in the analysis. A participant was considered active if they recorded at least 70% of the prescribed treatments. Data on patients with a current inhibitor or not on treatment with a factor concentrate were excluded. Being these time-dependent variables, the same

patient could still contribute with data on periods where these exclusion criteria did not apply.

Patients with an individual PK available (# 1 and 3 above) were also included in the analysis for the population without an individual PK available (# 2 and 4 above). To do so, the individual PK information was ignored when contributing to cohorts # 2 and 4.

**Outcome**

The outcome of the study was the number of bleeds occurring during the follow-up period. A bleed was defined as any treated bleeding episode, i.e., an event interpreted as a bleed by the patients or their physician and treated with factor concentrate. These events were registered by the patients. We did not count bleeds occurring within 72 hours of another bleed at the same site, assuming that those were follow-ups of the original bleeds.[4] Bleeds occurring in the setting of surgery were also removed.

**Predictors**

Based on the derivation phase, the only predictor included in the model was the factor concentration at a given moment. The concentration was calculated using PK data from the WAPPS-Hemo database and treatment data from CBDR. For PWH with no available individual PK, these were predicted from the typical values of Population PK models developed in the frame of the WAPPS-Hemo project. The mapping between the Population PK model and factor concentrate is reported in **sTable 1**.

The relationship between the hazard and time was described using a Gompertz model, and the relationship between the hazard and the factor concentration was described using a Hill model. The final model was the following:

$$h(t) = h_0(t+1)^{-\beta} \frac{FVIII/IX}{FVIII/IX + EC50} e^{\eta}$$

The parameters' estimates are reported in **Table 4-14**.

The following predictors were evaluated and not included in the final model: age, body weight, body mass index (BMI), treatment category (on-demand versus regular prophylaxis), and type of factor concentrate used (standard versus extended half-life). As reported in the supplementary material, this was based on the minimal contribution of these variables to explaining the between-subject variability (BSV) in the risk of bleeding.

**Power/sample size**

We included all eligible patients during the study period to maximize power and confidence in the results. A graph with the estimated power to show a slope different than 1 (the null hypothesis) for the regression of the predicted versus the observed number of bleeds for a variety of alternative slopes and sample sizes are reported in the supplementary material.

**Missing data**

Measures to minimize missing data are already in place in CBDR through educational campaigns and periodical data quality activities. As mentioned in the participants section, we excluded patients recording less than 70% of their prescribed treatments. Above that threshold, we assumed that reporting was complete, as we could not distinguish treatments not recorded from those that were missed. The methods used for dealing with missing data to derive the pop-PK profiles in WAPPS-Hemo have been described in the derivation manuscript (see the supplementary material). Fat-free mass is a known covariate of these Population PK models. Since fat-free mass is calculated from weight, height, and age, these

parameters were required. Handling of missing height was performed by assuming the median height obtained for the patient's age in the NHANES database.[5]

**Statistical analysis methods**

The baseline characteristics of the population were tabulated using standard descriptors of central tendency and variability (mean and standard deviation [SD] or median and first and third quartiles [Q1, Q3], as appropriate) for continuous variables and percentages for categorical variables. To assess the predictive performance of the model, we fed it with the recorded treatment data and PK parameters. We used this information to generate the predicted number of bleeds occurring during the available follow-up time for each patient. The predicted number of bleeds was the median cumulative hazard produced by the model.[6] The calibration of the risk score predictions was evaluated by plotting the predicted versus the observed number of bleeds in the follow-up period, and by calculating the slope and intercept of the regression line.[2] The regression model took into account for clustering of multiple treatment periods in patients. An intercept of zero and a slope of one would have indicated perfect calibration. We also checked what proportion of the observed number of bleeds fell within the $5^{th}$ and $95^{th}$ percentiles of the predicted (i.e., $5^{th}$ and $95^{th}$ percentile of the cumulative hazard). For this analysis, we rounded the predicted number of bleeds to zero if the cumulative hazard was <0.5, and to one if the cumulative hazard was 0.5-1. Furthermore, we visually compared the observed and predicted Kaplan-Mayer curves for the occurrence of the first bleed in each cohort. Moreover, we calculated the accuracy of the model for predicting the occurrence of an annualized bleeding rate (ABR) ≥4 bleeds/year. The ABR was calculated as follows: $number\ of\ bleeds \times \frac{365}{days\ of\ observation}$. To

obtain a stable estimate of the ABR, we only included in this analysis patients with a minimum follow-up of six months. The ABR threshold of 4 bleeds/year was set by consensus between the hemophilia treaters among the authors' group as the ABR at which they would consider changing the treatment for poor efficacy. Similarly, the minimum of 6 months of FUP to calculate the ABR was set by consensus among the authors. Finally, the analyses on the observed versus the predicted number of bleeds were replicated using the ABR as the outcome. For all these measures, 95% confidence intervals (CIs) and p-values were calculated. The criterion for statistical significance was set at $\alpha = 0.05$.

The PopPK analysis and the risk assessment model were obtained using non-linear mixed-effects modeling as implemented in NONMEM and PDxPop with the Laplacian estimation method (version 7.3 and version 5.2, respectively; ICON Development Solutions, Ellicott City, MD, USA). The risk estimates were performed using R (R Core Team 2021, v 3.6.1, R Foundation for Statistical Computing, Vienna, Austria). The descriptive and validation analyses were performed using STATA (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC).

**Ethics**

The WAPPS and CBDR projects have received approval from the Hamilton Integrated Research Ethics Board (HIREB). The PMCH study was also approved by HIREB, and the creation of a RAM was pre-specified as a secondary analysis of that study.

## Results

### Participants

One hundred forty-two PWH A had individual PK information and were included in the main analysis. The median (Q1; Q3) age was 31 (17; 44) years. These 142 participants were observed for 151 treatment periods of a median duration of 360 (278; 362) days. They were treated with extended half-life (EHL) products in 79 (52%) treatment periods. During the follow-up, the participants recorded 337 bleeds, with a median of 1 (0; 3) bleed per treatment period, and a median ABR of 1.0 (0.0; 3.1) bleed/year. No bleeds occurred in 54 (41.6%) treatment periods.

When ignoring the individual PK information and extending the analysis to participants without individual PK data, 483 PWH A were included. Their median age was 24 (11; 37) years, and they were observed for 632 treatment periods of a median duration of 362 (166; 1097) days. They were treated with extended half-life (EHL) products in 167 (26%) treatment periods. During the follow-up, the participants recorded 1998 bleeds, with a median of 1 (0; 4) bleed per treatment period, and a median ABR of 2.8 (1.0; 3.1) bleeds/year. Zero bleeds occurred in 162 (35.3%) treatment periods.

The characteristics of the population and outcomes distribution, including PWH B with and without an individual PK available, are reported in **Table 4-15**.

### Model performance

***PWH A and individual PK available.*** The graph with the observed versus the predicted number of bleeds during each treatment period for PWH A and an individual PK available is reported in **Figure 4-6**. The regression coefficient (95% CI) was 1.63 (-1.05; 4.31), and the intercept was 0.83 (-1.60; 3.27), with an $R^2$ of 0.02 (see **Table 4-16**). The observed number

of bleeds fell within the 5th and 95th percentiles of the prediction in 125/151 observation periods (82.8%, 95% CI 76.8%; 88.8%). The observed and predicted Kaplan-Mayer curves for the first bleed for PWH A and an individual PK available are reported in **Figure 4-7**. The graph with the observed versus predicted ABR and the results of the regression analysis are reported in the supplementary material. The model did not predict the occurrence of an ABR ≥4 bleeds/year in any of the PWH A and an individual PK available (data in the supplementary material).

***PWH A and no individual PK available.*** The graph with the observed versus the predicted number of bleeds during each treatment period for PWH A without an individual PK available (or obtained ignoring the individual PK information) is reported in **Figure 4-8**Figure 4-8: observed versus predicted number of bleeds during each treatment period, for PWH A without an individual PK available. The regression coefficient (96% CI) was 1.97 (1.35; 2.59), and the intercept was 0.45 (-0.56: 1.46), with an $R^2$ of 0.09 (see **Table 4-16**). The observed number of bleeds fell within the 5th and 95th percentiles of the prediction in 446/632 observation periods (70.6%, 95% CI 67.0%; 74.1%). The observed and predicted Kaplan-Mayer curves for the first bleed for PWH A without an individual PK available are reported in **Figure 4-9**. The graph with the observed versus predicted ABR and the results of the regression analysis are reported in the supplementary material. The model did not predict the occurrence of an ABR ≥4 bleeds/year in any of the PWH A without an individual PK available (data in the supplementary material).

The results on the model performance for PWH B with and without an individual PK available are similar and are reported in the supplementary material.

## Discussion

In our population, the number of observed bleeds was in the 95% range of the predictions in 82.8% of PWH A and an individual PK available, and less than 70% of the remaining patients. The model tended to overestimate the number of bleeds, particularly in the population with zero bleeds. This was more evident in PWH, with a point estimate for the intercept >2 in both people with and without an individual PK available. On the other hand, the model tended to underestimate the number of bleeds in participants with frequent bleeding episodes. The model predicted more than four bleeds per year, a threshold that our clinical experts suggested being clinically relevant, only in one patient (and this patient actually bled less). The model contributed to only partially explaining the variability in the outcome ($R^2$ ranging from 0.02 to 0.13 in the different study cohorts). While the model proposed by Chelle et Al. helped with confirming the association between plasmatic factor levels and bleeding risk, as reported in the supplementary material, its performance when used to predict the number of bleeds was less than ideal. This could be explained by the fact that the model predicts the bleeding risk only based on the plasmatic factor levels. It is known that the bleeding phenotype can vary widely in patients with similar baseline factor levels.[7] It is not surprising that the same applies to plasmatic factor levels after the administration of exogenous coagulation factor. In our systematic review of risk factors for bleeding in hemophilia,[1] we found that at least the bleeding history, physical activity levels, obesity, and concomitant antithrombotic treatment should be considered when building a RAM to predict the risk of bleeding in this population. Obesity was not included in this model because it was not found to significantly reduce the between-subject variability, as reported in the supplementary material. Unfortunately, no data were available for

bleeding history, physical activity levels, and antithrombotic treatment. Regarding the bleeding history, this was not available because all the available bleeding data were used as outcomes for the model, to maximize the power of the analysis. An alternative approach would have been to use part of the follow-up period as historical information. The fact that the model performance was worse for participants with zero bleeds and for frequent bleeders might be at least in part explained by the fact that the treatment history was not included in the model. Physical activity and concomitant antithrombotic treatment, instead, were not systematically collected in CBDR and therefore actually not available in the dataset. The model performance might also be partially explained by the fact that the model was primarily designed to explore the association between factor levels and risk of bleeding, more than to specifically predict the number of bleeds in PWH. Methodological considerations about how to derive and validate a RAM in this context will be discussed in the next chapter. A further limitation of the present study is that it relies on routinely collected data and bleeds and treatments are reported by patients. These data are likely less precise than data collected with the help of dedicated research personnel, and this might have negatively affected both the derivation and validation process and the performance of the model. Finally, one might argue that the ongoing development of innovative therapies that continues to push the boundaries of hemophilia management might eliminate the need for a risk assessment model based on the pharmacokinetics characteristics of factor concentrates. However, it is still unknown if these new therapies are as good as factor concentrates to achieve enough protection that abate the risk of bleeding in certain high-risk situations. Moreover, considering the costs of these new treatments, it is likely that

factor concentrates will continue to be used for the time being, at least in selected populations.

## Conclusions

The performance of the RAM assessed in this study does not allow us to use it for guiding decision-making on treatment strategies based on the predicted number of bleeds. The need for a RAM that can accurately predict the risk of bleeding in individual PWH A and B is still unmet.

## Funding

The PMCH study was funded with grants from the Canadian Institute of Health Research (CIHR) and the Shire - Canadian Hemophilia Epidemiological Research Program (S-CHERP). Federico Germini received financial support through the Physicians' Services Incorporated (PSI) Research Trainee Award.

## Authors' contribution

FG and AI conceived this study. FG is the guarantor of the study and drafted the manuscript. EI prepared the data. PC performed the PK analyses and the outcome predictions, FG the validation analysis. All the authors read, provided feedback, and approved the final manuscript.

## References

1 Germini F, Noronha N, Abraham Philip B, *et al.* Risk factors for bleeding in people living with hemophilia A and B treated with regular prophylaxis: A systematic review of the literature. *J Thromb Haemost* 2022;**20**:1364–75. doi:10.1111/jth.15723

2 Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;**162**:W1. doi:10.7326/M14-0698

3 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015;**12**:e1001885. doi:10.1371/journal.pmed.1001885

4 Abrantes JA, Solms A, Garmann D, *et al.* Relationship between factor VIII activity, bleeds and individual characteristics in severe hemophilia A patients. *Haematologica* 2020;**105**:1443–53. doi:10.3324/HAEMATOL.2019.217133

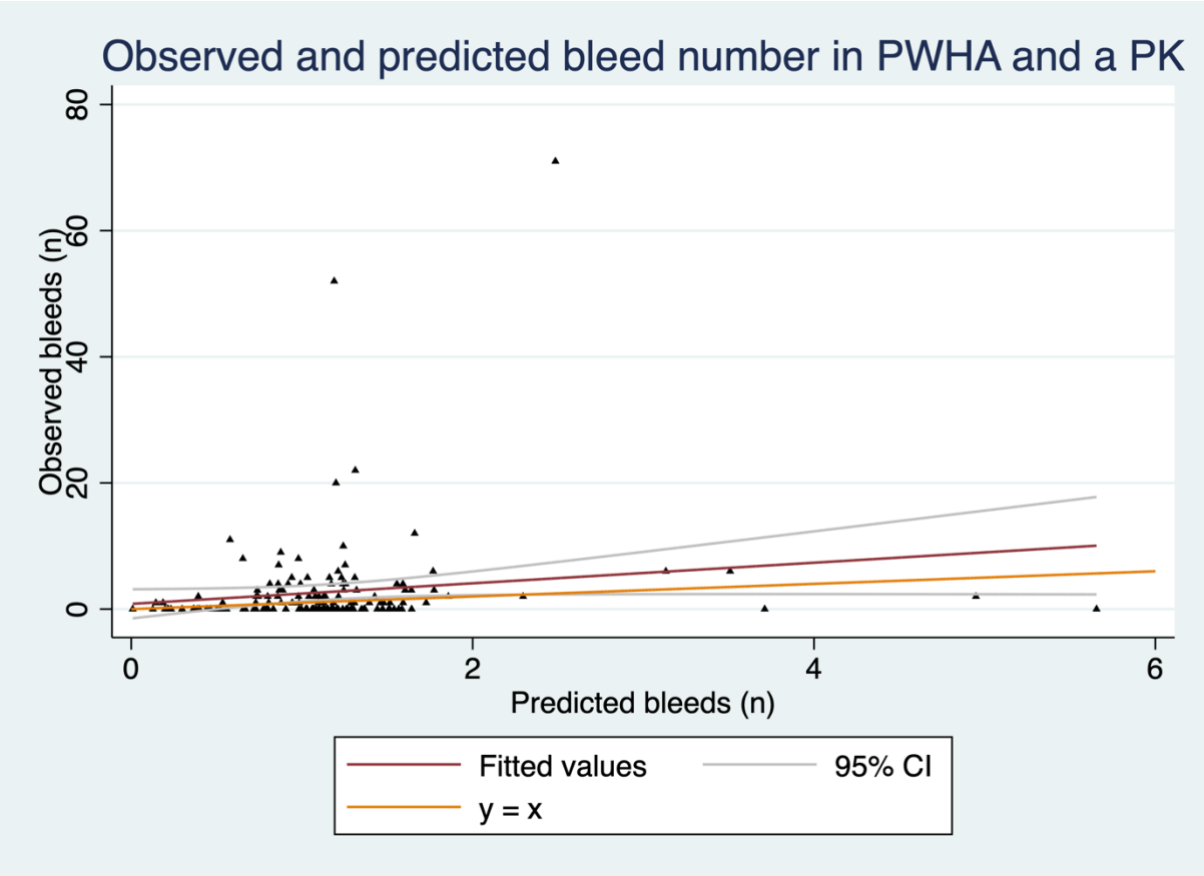
5 Fryar CD, Gu Q, Ogden CL, *et al.* Anthropometric Reference Data for Children and Adults: United States, 2011-2014. *Vital Health Stat 3* 2016;:1–46.

6 Clark TG, Bradburn MJ, Love SB, *et al.* Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 2003;**89**:232–8. doi:10.1038/SJ.BJC.6601118

7 den Uijl IEM, Mauser Bunschoten EP, Roosendaal G, *et al.* Clinical severity of haemophilia A: Does the classification of the 1950s still stand? *Haemophilia* 2011;**17**:849–53. doi:10.1111/j.1365-2516.2011.02539.x

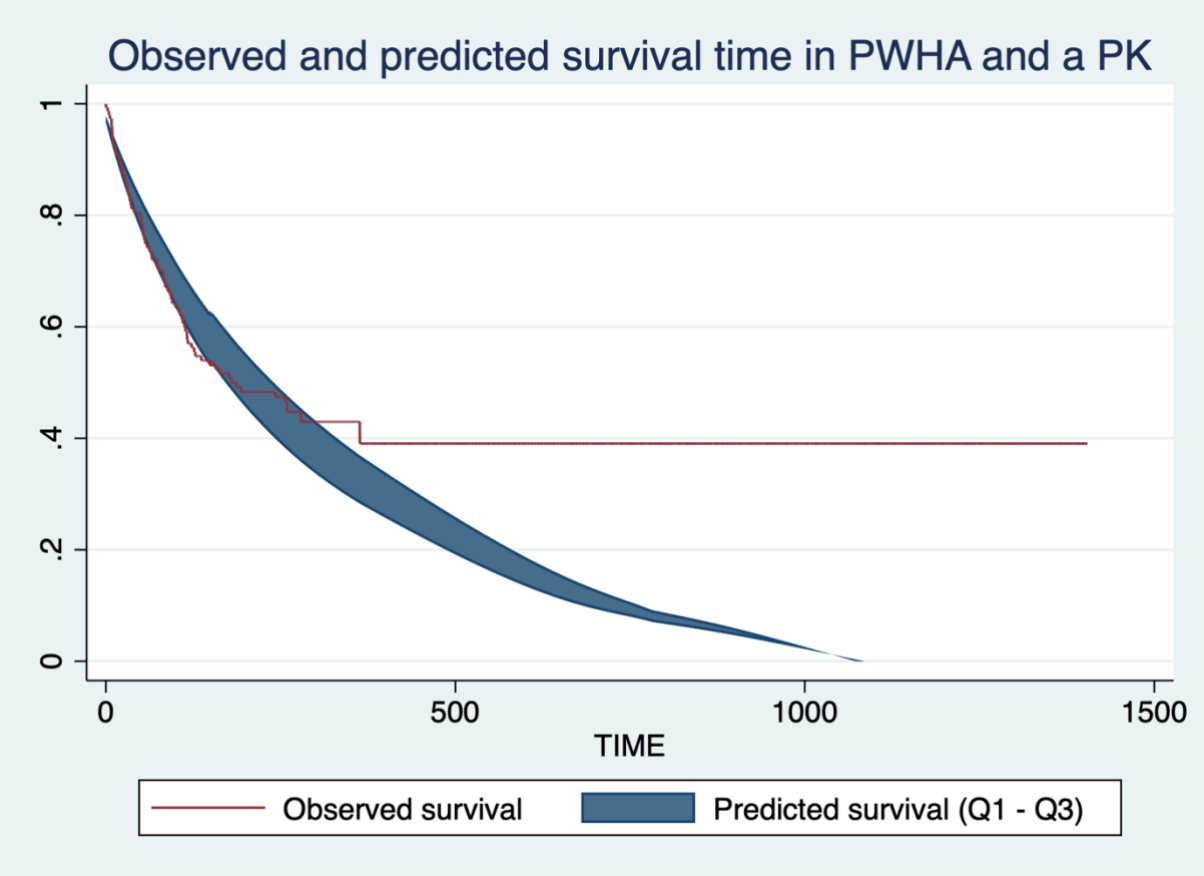
## Figures

*Figure 4-6: observed versus predicted number of bleeds during each treatment period, for PWH A and an individual PK available.*



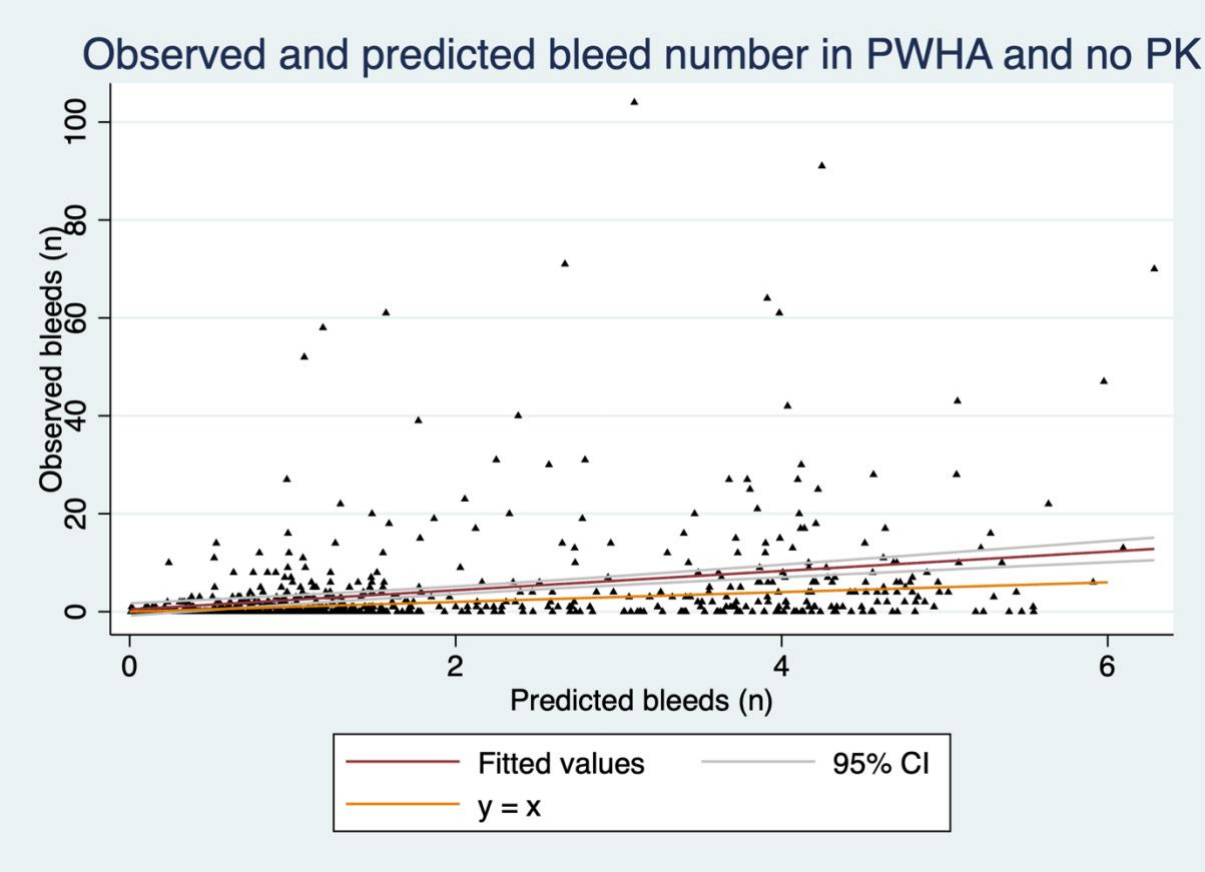

CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

*Figure 4-7: observed versus predicted survival time (first bleed) during each treatment period, for PWH A and an individual PK available.*



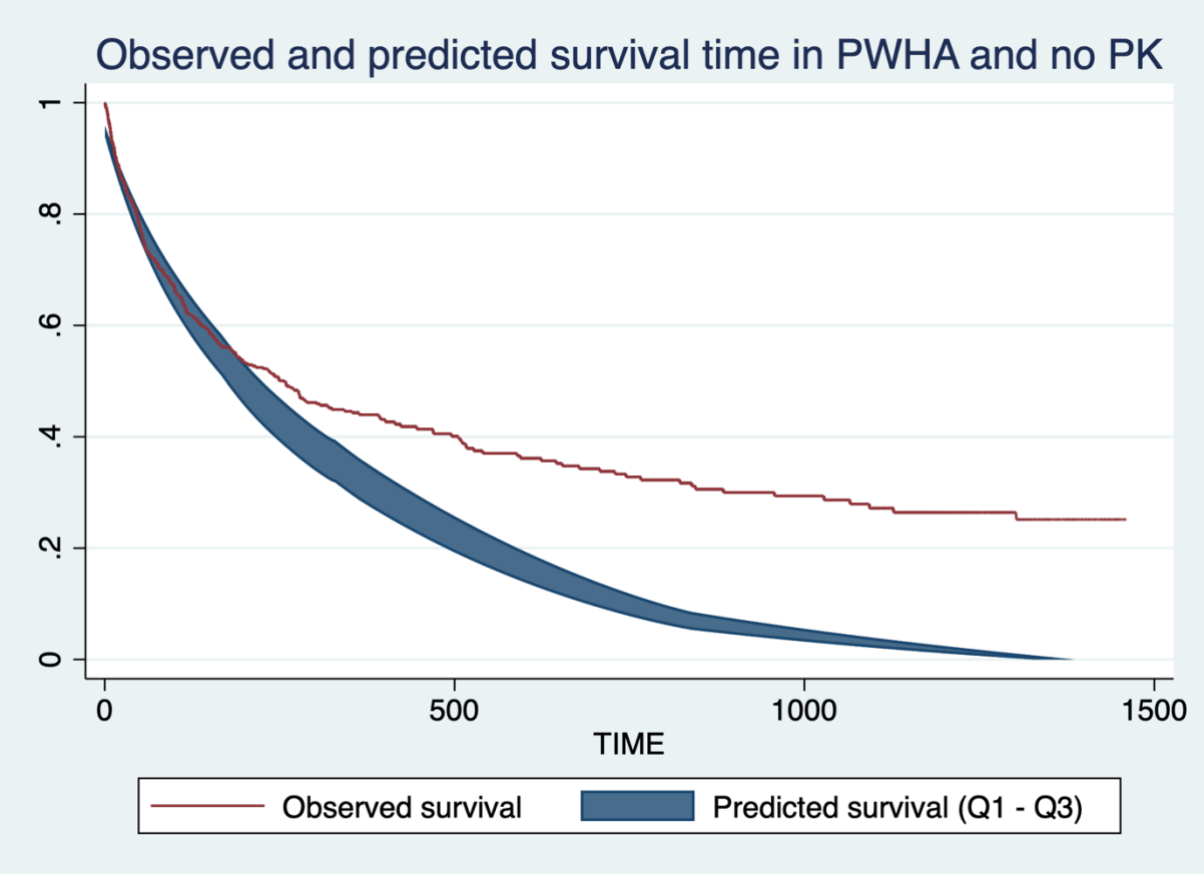

PK: pharmacokinetics; PWH: people with hemophilia; Q1: first quartile; Q3, third quartile.

*Figure 4-8: observed versus predicted number of bleeds during each treatment period, for PWH A without an individual PK available.*




CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

*Figure 4-9: observed versus predicted survival time (first bleed) during each treatment period, for PWH A without an individual PK available.*


Observed and predicted survival time in PWHA and no PK
0
.2
.4
.6
.8
1
0
500
1000
1500
TIME
Observed survival
Predicted survival (Q1 - Q3)

# Tables

*Table 4-14: Parameter estimates for the risk assessment model.*

| Parameter | Estimate |
|---|---|
| $h_0$ (bleeds/yr) | 6.51 |
| EC50 (IU/mL) | 0.140 |
| $\beta$ (1/yr) | -0.198 |
| η (%) | 94.9 |

EC50: factor activity resulting in half-maximum inhibition of the hazard, *η:* inter-individual deviation from the average hazard, expressed as *coefficient of variation (CV)*

*Table 4-15: Characteristics of the population and outcomes distribution.*

| | PWH A and PK | PWH B and PK | PWH A, no PK | PWH B, no PK |
|---|---|---|---|---|
| Participants (n) | 142 | 35 | 483 | 112 |
| Age (years) | 31 (17; 44) | 42 (18; 50) | 24 (11; 37) | 28 (13; 46) |
| Weight (kg) | 76 (63; 92) | 73 (65; 85) | 70 (48; 86) | 73 (51; 84) |
| FUP time (days) | 360 (328; 363) | 1142 (324; 1345) | 420 (359; 1446) | 1379 (798; 1455) |
| | | | | |
| Treatment periods (n) | 151 | 51 | 632 | 180 |
| SHL (n) | 72 (48%) | 4 (8%) | 465 (74%) | 84 (47%) |
| Dose (IU/kg) | 28.5 (24.7; 35.5) | 37.6 (27.2; 50.8) | 26.8 (20.1; 33.7) | 45.1 (36.6; 61.0) |
| Dose interval (days) | 2 (2; 3) | 3 (2; 3) | 2 (2; 3) | 3 (2; 4) |
| EHL (n) | 79 (52%) | 47 (92%) | 167 (26%) | 96 (53%) |
| Dose (IU/kg) | 36.2 (29.0; 43.5) | 50.3 (37.5; 64.9) | 37.6 (30.4; 45.5) | 45.8 (38.7; 57.7) |
| Dose interval (days) | 3 (2; 4) | 7 (6; 8) | 3 (3; 4) | 7 (6; 7) |
| FUP time (days) | 360 (278; 362) | 465 (193; 1124) | 362 (166; 1097) | 461 (167; 1181) |
| Total number of bleeds | 337 | 252 | 1998 | 637 |
| Median number of bleeds | 1 (0; 3) | 2 (0; 9) | 1 (0; 4) | 1 (0; 6) |
| Median ABR (bleeds/year)* | 1.0 (0.0; 3.1) | 2.3 (0.5; 5.0) | 2.8 (1.0; 3.1) | 1.1 (0.0; 4.0) |
| 0 bleeds, n* | 54 (41.6%) | 8 (19.5%) | 162 (35.3%) | 37 (28.0%) |

*Table 4-16: regression for observed versus predicted number of bleeds.*

| | PWHA with PK | PWHB with PK | PWHA no PK | PWHB no PK |
|---|---|---|---|---|
| n | 151 | 51 | 632 | 180 |
| Coefficient (95% CI) | 1.63 (-1.05; 4.31) | 2.83 (0.87; 4.79) | 1.97 (1.35; 2.59) | 1.57 (0.60; 2.53) |
| p-value | 0.230 | 0.006 | <0.001 | 0.002 |
| Intercept (95% CI) | 0.83 (-1.60; 3.27) | 2.39 (-0.78; 5.57) | 0.45 (-0.56: 1.46) | 2.16 (0.47; 3.84) |

| | | | | |
|---|---|---|---|---|
| p-value | 0.501 | 0.135 | 0.380 | 0.013 |
| $R^2$ | 0.02 | 0.13 | 0.09 | 0.07 |

CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

# Chapter 5 – Discussion, future directions, and conclusions.

In the previous chapters, we identified the risk factors that should be included in the derivation phase of a risk assessment model (RAM) for predicting the risk of bleeding in people living with hemophilia (PWH): plasmatic coagulation factor levels, bleeding history, physical activity, antithrombotic treatment, and obesity. We identified wrist-wearable devices with better acceptability and accuracy that could be used to collect data on physical activity, so far missing in our dataset. Finally, we validated the predictive performance of a model that allows estimating the bleeding risk based on the factor levels over time, derived using individual pop-PK data profiles and infusion logs. This model failed to accurately predict the number of bleeds in our cohort. We believe that this was due to issues with the data availability and the methodology used, which will need to be considered in the next step of this process, the model update.[1] In terms of data availability, efforts are needed to collect data on the risk factors identified in our systematic review that were missing in the database used to develop the RAM. A structured data collection on concomitant antithrombotic treatment could easily be added to the Canadian Bleeding Disorder Registry (CBDR). However, this should be followed by an active effort to insert the data from the patient or their care provider. A more efficient approach would be the integration of the registry with the Electronic Health Records (EHR) of their treatment center or the regional administrative databases where information about medication prescriptions is stored. This solution would require more computational effort initially, but data would then flow automatically. Data accuracy would depend on the availability of correct information on the above-mentioned databases. Unfortunately, the availability and accuracy of these data is

often suboptimal. One should promote some improvement, for example, through incentives to clinicians or using resources from research. This would generate a virtuous circle where high-quality data inputted into the EHR would facilitate clinical care and documentation, in turn allowing gathering of clean data that can secondarily be used for clinical research. However, the cumbersome absents in terms of data availability are physical activity and bleeding history. Data on physical activity could be collected manually (e.g. through questionnaires) or passively with the use of wearable devices. Again, the latter option is more appealing in terms of sparing human resources and long-term sustainability. This option would generate a large amount of data and requires careful thinking on how to store and analyze these data. The downside of this option is that people not using wearables for measuring physical activity would not be able to use the RAM. This might introduce some inequity. Data on treatment history are already available. As mentioned, in the past all the available data have been used to gather information about the outcomes, to increase the power of the analysis. With more data being collected over time, this could change. Another way of increasing the available data would be a collaborative effort with other bleeding disorder registries. This would allow dedicating some historical data (e.g. the first 6-12 months of treatment for each participant) to estimate the bleeding history. The rest of the data would be used as follow-up, while still retaining a large enough sample size. Such a combined effort is particularly needed for a rare disease like hemophilia, and there are examples of successful collaborations in the past, in this field.[2–5]

Regarding the methodology applied, we mentioned in Chapter 4 how the model we tried to use for our predictions was originally thought more to explore the association between plasmatic coagulation factor levels and bleeding risk than to accurately predict the risk for

bleeding in individuals. If we were doing this again, as we plan to do once we will have gathered the data described above, we could use different approaches to derive and validate a RAM.

One possibility would be would be to update the model sticking to classical statistics but using a different approach for some steps of the process, most importantly the variable selection and the internal validation. We will now describe these possible alternatives.

Regarding the variable selection, in our RAM, we decided not to include the body mass index (BMI). This was based on the finding that this variable did not contribute significantly to explaining the between-subject variability (BSV). If on one side this approach allows for deriving a leaner model, this may lid to overfitting.[6] An alternative approach is to force into the model variables that are deemed relevant by experts or, better, selected through a systematic review like the one described in Chapter 2.[1] The same will apply to the treatment history and physical activity, once these data will be available.

Regarding the internal validation process, this was initially conducted on the same sample used for the derivation, and on simulated patients, as described in the appendix to Chapter 4. When the model performance is assessed using this technique, it is called “apparent performance”, and it is described to lead to optimistic estimates, especially in small samples.[7] This is the reason why we decided to further validate the model, as reported in Chapter 4. It is not surprising that this translated into overfitting, with the model performance in the new sample being worse that the apparent performance in the original sample. The new data were coming from a similar population, being a sub-sample of the original one in terms of the clinical setting, but at a different time frame. This process, called temporal validation, has been described as “intermediate between internal validation and

external validation".[8] Due to these differences in the data source, it is normal to expect a worse model performance in the validation phase. However, this is the expression of optimistic estimates in the derivation and internal validation phases, and there are ways to take this into account when assessing the model performance. One example is the use of a technique called cross-validation. In cross-validation, the sample is divided into n sub-samples. The model is derived using data from n-1 sub-samples (e.g., nine out of ten). The model is then validated on the remaining sample (i.e. the tenth). The process is repeated n times, and the average model performance is calculated and reported.[1] Bootstrap validation is even more appealing, even though computationally more intense. In this case, after deriving the model and assessing its apparent performance, the process is repeated at least 100 times using new samples generated through sampling with replacement.[1] This allows estimating the optimism as the average of the difference between the apparent performance and the performance of each bootstrap model. This method also allows adjusting the model's regression coefficient using a shrinkage factor, so that the model performance in the external validation phase will improve.[9]

Another possible approach to the model derivation and validation is the use of machine learning.[10] Several methods could be used in this field. In the appendix, we report a protocol on the use of random forests[11] for this purpose, as an example. This was the final project of one of my Ph.D. courses, an independent study entitled "Machine Learning for Clinical Epidemiologists". Pros and cons of using machine learning for this purpose are discussed there.

We can conclude that a risk assessment model for the prediction of bleeding in people living with hemophilia should include plasmatic coagulation factor levels, bleeding history,

physical activity, antithrombotic treatment, and obesity. For measuring physical activity, we identified two devices from FitBit (Charge and Charge HR) as the most accurate for counting steps, and the Apple Watch as the most accurate for measuring heart rate. We ascertained that an existing risk assessment model with this aim provides optimistic estimates and identified possible causes for this suboptimal performance. This could be addressed by retaining key predictors in the model and using methods like bootstrapping to obtain a model that would perform better in an external validation phase and, therefore, be generalizable and used in everyday clinical activity.

## References


1 Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;**162**:W1–73. doi:10.7326/M14-0698

2 Iorio A, Stonebraker JS, Chambost H, *et al.* Establishing the prevalence and prevalence at birth of hemophilia in males a meta-analytic approach using national registries. *Ann Intern Med* 2019;**171**:542–6. doi:10.7326/M19-1208

3 Probestudy. https://probestudy.org/ (accessed 19 Mar 2023).

4 WAPPS-Hemo. https://www.wapps-hemo.org/ (accessed 19 Mar 2023).

5 Home | IPSG - International Prophylaxis Study Group. https://www.ipsg.ca/ (accessed 19 Mar 2023).

6 Steyerberg EW, Eijkemans MJC, Harrell Jr FE, *et al.* Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000;**19**:1059–79. doi:https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0

7 Steyerberg EW, Harrell FE, Borsboom GJJM, *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;**54**:774–81. doi:10.1016/S0895-4356(01)00341-9

8 Altman DG, Vergouwe Y, Royston P, *et al.* Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:1432–5. doi:10.1136/BMJ.B605

9 Steyerberg EW. *Clinical Prediction Models*. Cham: : Springer International Publishing 2019. doi:10.1007/978-3-030-16399-0

10 W L, D P, T T, *et al.* Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* 2016;**18**. doi:10.2196/JMIR.5870

11 Breiman L, Cutler A. Random forests — Classification description: Random forests. http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home. htm. 2007;:1–27.https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (accessed 11 Aug 2021).

# Supplementary material - Chapter 2

## Search Strategy

Search date: Aug. 21, 2019

Limits: Not animal - No date, age or language limit

Databases: OVID: MEDLINE &EMBASE. The Cochrane Central Register of Controlled Trials (CENTRAL) & the Cochrane Database of Systematic Reviews (CDSR).

| Database | Total Retrieved |
|---|---|
| Medline | 1435 |
| Embase | 1044 |
| CENTRAL | 25 |
| CDSR | 10 |
| Total | 2514 |
| After deduplication | 1858 |

| MEDLINE -- OVID Medline Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present | |
|---|---|
| hemophilia a/ or hemophilia b/ or (hemophilia or haemophilia).mp | 26017 |
| 2. exp Hemorrhage/ or Hemorrhage.mp. or haemorrhage.mp. or Hematemes*or haematemes*.mp. or Epitaxi*.mp. or subclinical bleeds.mp. or Hemarthrosi*.ti,ab. or haemarthrosi*.ti,ab. or pseudobleed.mp. or (bleed* adj2 (pattern* or hazard or frequency or probability or risk or rate or episode or window or phenotype or break-through or intra-articular or extra-articular or profile or traumatic or spontaneous or joint)).ti,ab. | 413841 |
| 3. Risk Factors/ or Risk Assessment/ or incidence.ti,ab. | 1549887 |
| 4. PROGNOSIS – SENS: incidence.sh. OR exp mortality OR follow-up studies.sh. OR prognos:.tw. OR predict:.tw. OR course:.tw. | 3275567 |
| 5. CPG – SENS: predict:.mp. OR scor:.tw. OR observ:mp. | 2265932 |
| 6. 3 or 4 or 5 | 4864246 |
| 7. 1 and 2 and 6 | 1454 |
| 8. Animals/ not (Animals/ and Humans/) | 4577069 |
| 9. 7 not 8 | 1435 |
| 10. 9 use ppez | 1435 |
| TARGET SET:<br>"23047359" [Unique Identifier]<br>"19143924" [Unique Identifier]<br>"11559935" [Unique Identifier]<br>"19822585" [Unique Identifier] | 6 |

| | |
|---|---|
| "28543946" [Unique Identifier]<br>"19298376" [Unique Identifier] | |

| Embase 1974 to 2019 April 22 | |
|---|---|
| 1. hemophilia a/ or hemophilia b/ or (hemophilia or haemophilia).mp | 39389 |
| 2. (exp bleeding/ or (Hemarthrosi* or haemarthrosi* or Hemorrhage or Haemorrhage or Hematemes* or haematemes* or pseudobleed).ti,ab. or (bleed* adj2 (pattern* or hazard or frequency or probability or rate or episode or window or break-through or intra-articular or extra-articular or profile or traumatic or spontaneous or joint or phenotype or risk)).ti,ab.) and (risk factor.mp. or risk factor/ or risk assessment/ or risk assessment models.mp.) | 85459 |
| 3. Prog – sens: exp disease course/ or risk:.mp. or diagnos:.mp. or follow-up.mp. or ep.fs. or outcome.tw. | 11746686 |
| 4. CPG – sens: predict:.tw. or exp methodology/ or validat:.tw. | 7335033 |
| 5. 3 or 4 | 15314613 |
| 6. 1 and 2 and 5 | 11606 |
| 7. (exp animal/ or nonhuman/) not exp human/ | 6183263 |
| 8. 6 not 7 | 1044 |
| 10. 9 use oemezd | |

| **CENTRAL -** Cochrane Library - https://www.cochranelibrary.com/advanced-search?cookiesEnabled | |
|---|---|
| #1 MeSH descriptor: [Hemophilia A] explode all trees | 360 |
| #2 MeSH descriptor: [Hemophilia B] explode all trees | 101 |
| #3 (hemphilia or haemophilia):ti,ab,kw | 1382 |
| #4 #1 or #2 or #3 1382 | 1382 |
| #5 MeSH descriptor: [Risk Factors] explode all trees | 24016 |
| #6 MeSH descriptor: [Risk Assessment] explode all trees | 8525 |
| #7 MeSH descriptor: [Hemorrhage] explode all trees | 13240 |
| #8 bleed* and (pattern* or hazard or frequency or probability or rate or episode or window or break-through or intra-articular or extra-articular or profile or traumatic or spontaneous or joint or phenotype or risk) | 25290 |
| #9 #5 or #6 | 29580 |
| #10 #8 or #9 | 53459 |
| #11 #4 and #9 and #10 | 9 |

| **CENTRAL -** Cochrane Library - https://www.cochranelibrary.com/advanced-search?cookiesEnabled |
|---|

| | | |
|---|---|---|
| #1 | MeSH descriptor: [Hemophilia A] explode all trees | 360 |
| #2 | MeSH descriptor: [Hemophilia B] explode all trees | 101 |
| #3 | (hemphilia or haemophilia):ti,ab,kw | 1382 |
| #4 | #1 or #2 or #3 | 1382 |
| #5 | MeSH descriptor: [Risk Factors] explode all trees | 24016 |
| #6 | MeSH descriptor: [Risk Assessment] explode all trees | 8525 |
| #7 | bleed* and (pattern* or hazard or frequency or probability or profile or phenotype or risk) | 18040 |
| #8 | MeSH descriptor: [Hemorrhage] explode all trees | 25290 |
| #9 | #5 or #6 or #7 | 46209 |
| #10 | #4 and #9 and #8 | |
| Cochrane Reviews = 10 | | |
| CENTRAL = 25 | | |

# Supplementary material – Chapter 3

## Database

The database for this systematic review is available on the journal's website, at this link.

## Outcome definitions:

Mean absolute percentage error (MAPE)

$$Mean\ absolute\ percentage\ error\ (MAPE) = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

n = number of times the summation iteration happens

$A_t$ = actual value (measured with the reference standard)

$F_t$ = Forecast value (measured with the index test)

Mean percentage error

$$Mean\ percentage\ error = \frac{1}{n}\sum_{t=1}^{n}\frac{A_t - F_t}{A_t}$$

n = number of times the summation iteration happens

$A_t$ = actual value (measured with the reference standard)

$F_t$ = Forecast value (measured with the index test)

Mean difference

$$Mean\ difference = \frac{1}{n}\sum_{t=1}^{n} A_t - F_t$$

n = number of times the summation iteration happens

$A_t$ = actual value (measured with the reference standard)

$F_t$ = Forecast value (measured with the index test)

Mean bias (Bland-Altman)

$$Mean\ bias\ (Bland-Altman) = \frac{1}{n}\sum_{t=1}^{n}(F_t - \frac{(A_t + F_t)}{2})$$

n = number of times the summation iteration happens

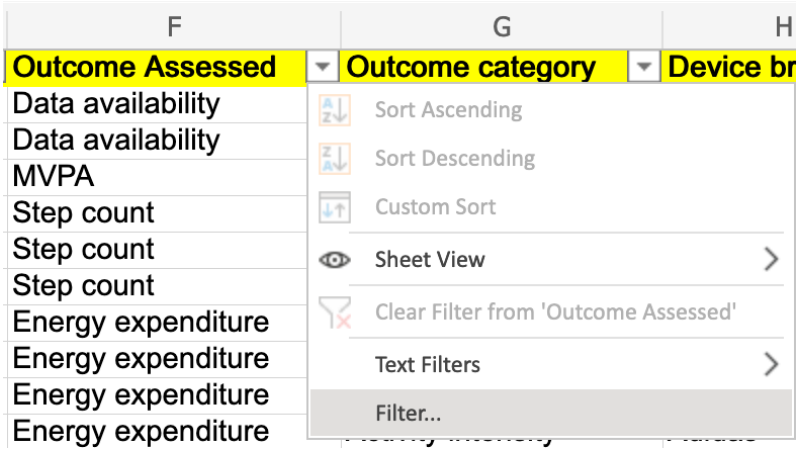
$A_t$ = actual value (measured with the reference standard)

$F_t$ = Forecast value (measured with the index test)

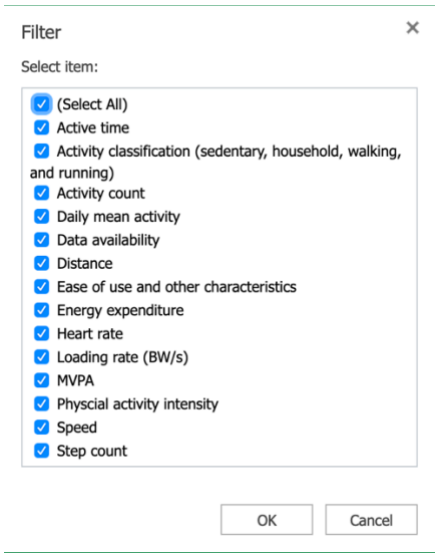
Supplementary Table 1: characteristics and results of all the studies included in the review (including the risk of bias assessment for each outcome).

The table was added as an excel file to the supplementary material. The reader can use the filter function at their preference, for example to isolate the objective (accuracy versus acceptability) the outcome within the objective (step count, energy expenditure, …), the device used as the index test, and the reference standard. An example is shown in **Supplementary figures 1 and 2**.

Supplementary Figure 1: accessing the filter function




Supplementary Figure 2: filters available for the variable “Outcome Assessed”


Filter
Select item:
(Select All)
Active time
Activity classification (sedentary, household, walking, and running)
Activity count
Daily mean activity
Data availability
Distance
Ease of use and other characteristics
Energy expenditure
Heart rate
Loading rate (BW/s)
MVPA
Physcial activity intensity
Speed
Step count
OK
Cancel

## Characteristics of the studies reporting on accuracy.

| Author, year | Device brand | Device model | Reference standard | Results | Scale of Measure |
|---|---|---|---|---|---|
| Outcome: Active time | | | | | |
| Hernandez-Vicente, 2016 | Polar | V800 | ActiTrainer | Mean (SD) bias (Bland-Altman) 32.0 (52.0) min, with mean (SD) 303.95 (93.29) min measured with the reference standard | Activity over 7-days under everyday conditions |
| Outcome: Activity classification (sedentary, household, walking, and running) | | | | | |
| Zhang, 2012 | GENE | Activ | Probably direct observation | For different machine learning algorithms (Logistic Regression, Decision Tree, Support Vector Machine, and Bayesian Network) the Incorrect classification rate ranged from 2.71 to 4.44% | 10-12 semi structured activities in lab or outdoor environment while wearing device |
| Outcome: Activity count | | | | | |
| Gironda, 2007 | Actiwatch | Score | VICON Motion Analysis System | Pearson's correlation coefficient 0.67-0.88 | Performance on two 15-minute trials of exercise activity prescribed for back-pain rehab. |
| Lawinger, 2015 | ActiGraph | GT3X+ | Manual count (video recording) | Correlation r .93, P < .001 "every 4000-vector-magnitude physical activity counts equal 27 arm motions". This 4000 was not pre-specified | Performance on 3 series of tasks: activities of daily living, rehab exercises and passive shoulder range at 5 specified velocities in one lab session. |
| Bruder, 2018 | ActivPAL | ActivPAL | 10-camera 3-D Motion analysis system (Vicon-MX3) | Mean difference -40.9 to 30.4 for different activities (95% CI reported) | Performance on two upper limb activities on week apart |
| Outcome: Daily mean activity | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Scott, 2017 | GENE | Activ | ActiGraph GT3X+ | Pearson's r 0.88 (95% CI = 0.82–0.93; p = <0.001) | Activity over 7-days under everyday conditions |
| Outcome: Distance | | | | | |
| Gaz, 2018 | Fitbit | Charge HR | Measured distance | Mean (SD) difference 0.028 (0.045) to 0.152 (0.124) m | Performance on a free walking or treadmill walking condition. Treadmill walking had pre-determined speeds. |
| Gaz, 2018 | Apple | Watch, series not NA | Measured distance | Mean (SD) difference 0.016 (0.05)to 0.037 (0.108) | |
| Gaz, 2018 | Garmin | Vivofit 2 | Measured distance | Mean (SD) difference 0.016 (0.028) to 0.107 (0.066) | |
| Gaz, 2018 | Jawbone | UP2 | Measured distance | Mean (SD) difference 0.008 (0.049) to 0.086 (0.059) | |
| Huang, 2016 | Jawbone | Up24 | Measured distance | Mean (SD) percentage error 5.2 (9.8) during flat ground walking (400 m) | Performance on slow, moderate, and fast walking speeds on treadmill Performance on slow, moderate, and fast walking speeds on treadmill |
| Huang, 2016 | Garmin | Vivofit | Measured distance | Mean (SD) percentage error 5.1 (11.4) during flat ground walking (400 m) | |
| Huang, 2016 | Fitbit | Flex | Measured distance | Mean (SD) percentage error -12.8 (15.4)% during flat ground walking (400 m) | |
| Wahl, 2017 | Beurer | AS80 | Measured distance | MAPE -51.9 to -17.6% | Performance on a treadmill for four 5 minute stages of different velocities, a 5-minute period of intermittent velocity, and a 2.4 km outdoor run and a 2.4 km outdoor run |
| Wahl, 2017 | Fitbit | Charge HR | Measured distance | MAPE -29.5 to -13.1% | |
| Wahl, 2017 | Fitbit | Charge | Measured distance | MAPE -29.9 to 16% | |
| Wahl, 2017 | Garmin | Vivofit | Measured distance | MAPE -25.0 to 23.3% | |
| Wahl, 2017 | Garmin | Vivosmart | Measured distance | MAPE -8.1 to 53.5% | |
| Wahl, 2017 | Garmin | Vivoactive | Measured distance | MAPE -6.1 to 51.4% | |
| Wahl, 2017 | Garmin | Forerunner 920XT | Measured distance | MAPE -3.3 to 26.0% | |

| | | | | | |
|---|---|---|---|---|---|
| Wahl, 2017 | Xaomi | Mi Band | Measured distance | Not Applicable (too many missing data, not analyzed) | |
| Wahl, 2017 | Withings | Pulse | Measured distance | MAPE 0.7 to 58.3% | |
| Outcome: Energy expenditure | | | | | |
| Stackpool, 2013 | Jawbone | UP | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.20 to 0.87 | First session completed on a treadmill at walking or running speed, selected by the participant. Second session was on elliptical cross-trainer at self-selected speed. Apart of the second session also took place in a gymnasium, where they completed agility ladder drills, basketball throws, and basketball lay-ups. |
| Stackpool, 2013 | Nike | Fuelband | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.08 to 0.72 | |
| Stackpool, 2013 | Fitbit | Ultra | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.24 to 0.67 | |
| Stackpool, 2013 | Adidas | MiCoach | Indirect calorimetry (Portable metabolic analyzer) | Pearson's r 0.55 to 0.81 | |
| Compagnat, 2018 | ActiGraph | GT3X+ | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Mean percentage difference 3% for walking subjects, 47% for subjects with wheelchair | Performed four tasks: transfers, manual tasks, walking on flat ground and walking up and down stairs. |
| Hargens, 2017 | Fitbit | Charge | ActiGraph GT3x | MAPE 30.6% | Activity over 7-days under everyday conditions |
| Mandigout, 2017 | Actical | Actical | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Spearman's r –0.19 (p 0.35) if weared on the plegic side, –0.27 (p 0.23) on the non-plegic side | Performance in various everyday tasks (transfer, walking, etc) within a laboratory setting |
| Mandigout, 2017 | ActiGraph | GTX | Indirect calorimetry (portable gas analyser, Metamax 3B, Cortex) | Spearman's r 0.08 (p 0.71) if wore on the plegic side, 0.20 (p 0.34) on the non-plegic side | |
| Montoye, 2017 | Fitbit | Charge HR | Indirect calorimetry (Parvo metabolic analyzer) | MAPE (SD) 43.7 (3.4) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, walking various speed and inclines, jogging, and cycling) |
| Price, 2017 | Fitbit | One | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias 2.91 (4.35) kcals/min | Walking on a treadmill at varying speeds |

| | | | | | |
|---|---|---|---|---|---|
| Price, 2017 | Garmin | Vivofit | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias -1.56 (2.34) kcals/min | |
| Price, 2017 | Jawbone | UP | Indirect calorimetry using ParvoMedics TrueOne 2400 | Mean (SD) bias 18.57 (30.17) kcals/min | |
| Roos, 2017 | Suunto | Ambi | Indirect calorimetry | MAPE 21.32 to 41.93% | Aerobic and anaerobic running on a treadmill in a laboratory setting |
| Roos, 2017 | Garmin | Forerunner 920XT | Indirect calorimetry | MAPE 11.54 to 49.30% | |
| Roos, 2017 | Polar | V800 | Indirect calorimetry | MAPE 10.1 to 39.5% | |
| Alsubheen, 2016 | Garmin | Vivofit | Indirect Calorimetry (Sable Systems International, Las Vegas NV) | Systematically underestimated by 29.5% during treatdmill walking test, p | Performance on treadmill walking tasks, and office activities within a laboratory session, completed in separate sessions on different days |
| Boeselt, 2016 | Polar | A300 | BodyMedia SenseWear | Pearson's r 0.74 ($p < 0.01$) | Performance in everyday conditions |
| Choi, 2010 | ActiGraph | GT1M | Room calorimeter | Mean (SD) percentage difference: 0.5 (8.0)% | Monitored through a 24-h stay in a laboratory setting. Stay included light activities, eating, sleeping, and participants were encouraged to complete normal day activities during downtime. |
| Chowdhury, 2017 | Microsoft | Band | CamNtech Actiheart | MAPE (SD) 34 (10)% | Performance against criterion measurements in both controlled laboratory conditions (simulated activities of daily living and structured exercise) and over a 24-hour period in free-living conditions. |
| Chowdhury, 2017 | Apple | Watch, series not NA | CamNtech Actiheart | MAPE (SD) 15 (10)% | |
| Chowdhury, 2017 | Jawbone | Up24 | CamNtech Actiheart | MAPE (SD) 30 (11)% | |
| Chowdhury, 2017 | Fitbit | Charge | CamNtech Actiheart | MAPE (SD) 16 (8)% | |
| Chowdhury, 2017 | Microsoft | Band | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 40 (16)% | |
| Chowdhury, 2017 | Apple | Watch, series not NA | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 27 (19)% | |

| | | | | | |
|---|---|---|---|---|---|
| Chowdhury, 2017 | Jawbone | Up24 | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 36 (14)% | |
| Chowdhury, 2017 | Fitbit | Charge | Indirect calorimetry (portable gas analyser, COSMED K4b2) | MAPE (SD) 36 (22)% | |
| Dondzila, 2018 | Fitbit | Charge HR | MET values of treadmill intensities | MAPE -8.4 to 89.2% | Performance on four-five minute stage treadmill tasks in a laboratory session and later in free-living conditions for one day. |
| Dondzila, 2018 | Miio | FUSE | MET values of treadmill intensities | MAPE 0 to 44.9% | |
| Durkalec-Michalski, 2013 | ActiGraph | GT1M | Indirect Calorimetery | Overestimated EE at moderate intensity by 60% and underestimated EE by 40% at vigorous intensity. 86% accurate in measuring EE at light intensity in relation to the values measured by indirect calorimetry. | Performance on leisure and exercise activities at various intensities in laboratory and free-living conditions |
| Ferguson, 2015 | Misfit | Shine | BodyMedia SenseWear | Mean absolute difference 468, mean (SD) measured with the reference standard = 3005 (569) | Activity under free-living conditions over 48 hours |
| Ferguson, 2015 | Jawbone | UP | BodyMedia SenseWear | Mean absolute difference 866, mean (SD) with the reference standard = 3005 (569) | |
| Hernandez-Vicente, 2016 | Polar | V800 | Actigraph ActiTrainer | Mean (SD) bias (Bland-Altman) 957.5 (679.9) kcal, with mean (SD) 1,456.48 (731.40) kcals measured with the reference standard | Activity under free-living conditions over 7 days |
| Lemmens, 2018 | Phillips | Optical Heart Rate monitor | Indirect calorimetry (portable gas analyser, COSMED K4b2) | Mean percentage error -2.6% | Performance on paced and self-paced exercise activities as well as household activities under laboratory conditions |
| Sirard - Phase 2 (Lab), 2017 | Movband | Movband | Indirect calorimetry system (Oxycon Mobile, Carefusion, Inc.) | Spearman's r 0.61 | Performance on structured activities (sitting, self-paced walking, catch, tag, |

| | | | | | |
|---|---|---|---|---|---|
| Sirard - Phase 2 (Lab), 2017 | Sqord | Sqord | Indirect calorimetry system (Oxycon Mobile, Carefusion, Inc.) | Spearman's r 0.87 | jogging) within a laboratory condition over 2 days |
| Wallen, 2016 | Apple | Watch, series not NA | Indirect calorimetry (MetaMax 3B, Cortex, Germany) | Mean (SD) bias (Bland-Altman) -123.1 (55.6) kcal, with index test mean (SD) = 285.7 (50.2) | Completed ~1-hr protocols involving supine and seated rest, walking and running on a treadmill and cycling on an ergometer in a laboratory condition |
| Wallen, 2016 | Fitbit | Charge | Indirect calorimetry | Mean (SD) bias (equation reported, since Bland-Altman parameters were systematically biased) 0.61*mean–224.6 (59.1) kcal, with index test mean (SD) = 236.8 (77.0) | |
| Wallen, 2016 | Samsung | Gear S | Indirect calorimetry | Mean (SD) bias (Bland-Altman) -26.1 (24.2) kcal, with index test mean (SD) = 261.4 (47.5) | |
| Wallen, 2016 | Miio | Mio alpha | Indirect calorimetry | Mean (SD) bias (equation reported, since Bland-Altman parameters were systematically biased) 0.91*mean - 318.77 (84.8) kcal, with index test mean (SD) = 236.8 (77.0) | |
| Woodman, 2017 | Garmin | Vivofit | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 44.6 (~8) | Completed 11 activities ranging from sedentary behaviors to vigorous intensities in a laboratory condition over one day |
| Woodman, 2017 | Withings | Pulse | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 63.7 (~4.5) | |
| Woodman, 2017 | Basis | Peak | Indirect calorimetry (Oxycon Mobile, Carefusion, Inc.) | MAPE (SD) 27.2 (~20) | |
| Imboden, 2018 | Fitbit | Flex | Indirect calorimetry | Mean percentage bias = -13% | Participated in an 80-minute protocol of exercises in a laboratory condition |
| Imboden, 2018 | Jawbone | Up24 | Indirect calorimetry | Mean percentage bias = -26% | |
| Wahl, 2017 | Polar | Loop | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE 5.6 to 56.4% | Performed a running protocol consisting of four 5 min stages of different constant velocities (4.3; 7.2; 10.1; 13.0 km·h−1), a 5 min period of intermittent velocity, and a 2.4 km outdoor run (10.1 km·h−1). |

| | | | | | |
|---|---|---|---|---|---|
| Wahl, 2017 | Beurer | AS80 | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -48.4 to 17% | |
| Wahl, 2017 | Fitbit | Charge HR | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -12.0 to 83.3% | |
| Wahl, 2017 | Fitbit | Charge | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -4.5 to 75.0% | |
| Wahl, 2017 | Bodymedia | Sensewear | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -25.3 to -1.4% | |
| Wahl, 2017 | Garmin | Vivofit | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -21.3 to 18.7% | |
| Wahl, 2017 | Garmin | Vivosmart | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -1.5 to -35.8% | |
| Wahl, 2017 | Garmin | Vivoactive | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX | MAPE -4.5 to 36.8% | |

| | | | | | |
|---|---|---|---|---|---|
| | | | Biophysik GmbH, Leipzig, Germany) | | |
| Wahl, 2017 | Garmin | Forerunner 920XT | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -26.6 to -9.2% | |
| Wahl, 2017 | Xaomi | Mi Band | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | Not applicable (too many missing data, not analyzed) | |
| Wahl, 2017 | Withings | Pulse | Indirect calorimetry with portable gas analyzer Metamax 3B (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) | MAPE -38.9 to -16.9% | |
| Dooley, 2017 | Apple | Watch, series not NA | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 16.54 (~13) to 210.84 (~96)% | Participants completed a 10-minute seated baseline assessment; separate 4-minute stages of light-, moderate-, and vigorous-intensity treadmill exercises; and a 10-minute seated recovery period in a laboratory setting |
| Dooley, 2017 | Fitbit | Charge HR | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 16.85 (~14) to 84.98 (~46)% | |
| Dooley, 2017 | Garmin | Forerunner 225 | Indirect calorimetry with Parvo Medics TrueOne 2400 (Parvo Medics Inc, Sandy, UT, USA) | MAPE (SD) 30.77 (~26) to 155.05 (~164)% | |
| Outcome: Heart rate | | | | | |
| Jo, 2016 | Basis | Peak K | Standard 12- lead electrocardiograph system (Cosmed C12x; Concord, CA, USA) | Mean(SD) bias (Bland-Altman) -3 (11) bpm | Each participant completed an initial rest period of 15 minutes followed by 5-minute periods of each of the following activities: 60W and 120W cycling, walking, |

| | | | | | |
|---|---|---|---|---|---|
| Jo, 2016 | Fitbit | Charge | Standard 12- lead electrocardiograph system (Cosmed C12x; Concord, CA, USA) | Mean(SD) bias (Bland-Altman) -9 (17) bpm | jogging, running, resisted arm raises, resisted lunges, and isometric plank. In between each exercise task was a 5-minute rest period. |
| Montoye, 2017 | Fitbit | Charge HR | Nonin PureSAT Pulse Oximeter | MAPE (SD) 6.6 (0.6) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, walking various speed and inclines, jogging, and cycling) |
| Dondzila, 2018 | Fitbit | Charge HR | Polar heart rate monitor | Trend to report lower mean heart rate values at running speeds of 134.1 m·min-1 and 160.9 m·min-1, compared to the Polar. | Performance on four-five minute stage treadmill tasks in a laboratory session and later in free-living conditions for one day. |
| Dondzila, 2018 | Miio | FUSE | Polar heart rate monitor | Mean heart rate values within 1.1 beats·min-1 of the Polar. | |
| Gillinov, 2017 | Garmin | Forerunner 235 | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 4.6 (7.7) to 13.7 (16.8)% | Completed exercise protocols on a treadmill, a stationary bicycle, and an elliptical trainer (Tarm movement) in a laboratory setting over one day. |
| Gillinov, 2017 | TomTom | Spark | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 4.5 (5.3) to 6.7 (9.6)% | |
| Gillinov, 2017 | Apple | Watch, series not NA | ECG leads, polar H7 chest strap monitor, | MAPE (SD) 3.2 (4.9) to 6.5 (10.8)% | |
| Gillinov, 2017 | Fitbit | Blaze | ECG leads, polar H7 chest strap monitor, scosche rhythm on forearm | MAPE (SD) 5.6 (6.4) to 15.9 (18.2)% | |
| Powierza, 2017 | Fitbit | Charge | Electrocardiogram | Mean(SD) bias (Bland-Altman) -6.04 (10.40) bpm | Completed the Buffalo Concussion Treadmill Test in a laboratory setting over one day. |

| | | | | | |
|---|---|---|---|---|---|
| Støve, 2019 | Garmin | Forerunner | Polar device | Mean difference (SD) 1 (2.3) to 17 (13.36) bpm, with mean (SD) frequence ranging from 59.5 (10.8) to 165.8 (16.4) | Performance during rest and three exercise conditions at submaximal level including cycling, treadmill, walking, running and rapid arm movement in a laboratory setting |
| Thomson, 2019 | Apple | Watch, series not NA | ECG | Mean percentage error 2.4 to 5.1% | Measured over performance in different intensity levels of activity, from very light to very rigorous, in a laboratory session over 1 day. |
| Thomson, 2019 | Fitbit | Charge HR 2 | Electrocardiogram | Mean percentage error 3.9 to 13.5% | |
| Wallen, 2016 | Apple | Watch, series not NA | ECG | Mean (SD) bias (Bland-Altman) -1.3 (4.4) bpm, with index test mean (SD) = 102.0 (14.4) | Completed ~1-hr protocols involving supine and seated rest, walking and running on a treadmill and cycling on an ergometer in a laboratory condition |
| Wallen, 2016 | Fitbit | Charge | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -9.3 (8.5) bpm, with index test mean (SD) = 102.0 (14.5) | |
| Wallen, 2016 | Samsung | Gear S | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -7.1 (10.3) bpm, with index test mean (SD) = 100.5 (14.6) | |
| Wallen, 2016 | Miio | Mio alpha | Electrocardiogram and indirect calorimetry | Mean (SD) bias (Bland-Altman) -4.3 (7.2) bpm, with index test mean (SD) = 102.0 (14.4) | |
| Dooley, 2017 | Apple | Watch, series not NA | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) 1.4 (~1) to 6.7 (~11)% | Participants completed a 10-minute seated baseline assessment; separate 4-minute stages of light-, moderate-, and vigorous-intensity treadmill exercises; and a 10-minute seated recovery period in a laboratory setting |
| Dooley, 2017 | Fitbit | Charge HR | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) 2.4 (~1.5) to 17.0 (~20.0) | |
| Dooley, 2017 | Garmin | Forerunner 225 | ActiGraph GT3X+, Polar Heart Rate Monitor, Pravo Medica TrueOne 2400 | MAPE (SD) ranging from 7.8 (~17) to 24.38 (~26) | |

| | | | | | |
|---|---|---|---|---|---|
| Stiles, 2013 | GENE | Activ | Advanced Mechanical Technology Inc. force plate | Sensitivity 97.6%, specificity 75.0%, overall agreement 85.6%, using a cut-off point of 3.125 g | Performed walking (slow, fast, and with bag), floor sweeping, running (slow and fast), jumping (low, G5 cm; high, 95 cm), and box drop (20 cm) in a laboratory session. |
| Stiles, 2013 | ActiGraph | GT3X+ | Advanced Mechanical Technology Inc. force plate | Sensitivity 90.5%, specificity 81.3%, overall agreement 85.6 using a pre-specified cut-off cut-off point 2.840 g | |
| Outcome: MVPA | | | | | |
| Semanik, 2020 | Fitbit | Flex | ActiGraph GT3X | Mean (SD) difference 18.5 (11.3), with mean (SD) 239.5 (86.2) min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Hargens, 2017 | Fitbit | Charge | ActiGraph GT3x | MAPE 46.3% | Activity over 7-days under everyday conditions |
| Scott, 2017 | GENE | Activ | ActiGraph GT3X+ | Pearson's r 0.84 (95% CI = 0.77–0.89; p <0.001) | Activity over 7-days under everyday conditions |
| Boeselt, 2016 | Polar | A300 | Bodymedia-SenseWear (SWA) device | Pearson's r -0.25 (p <0.01) | Performance in everyday conditions |
| Ferguson, 2015 | Misfit | Shine | Actigraph GT3X+ | Mean absolute difference (MAD) = 15.2, mean (SD) with the reference standard = 58.5 (37.6) | Activity under free-living conditions over 48 hours |
| Ferguson, 2015 | Jawbone | UP | Actigraph GT3X+ | Mean absolute difference (MAD) = 18.0, mean (SD) with the reference standard = 58.5 (37.6) | |
| Redenius, 2019 | Fitbit | Flex | Actigraph GT3X+ | MAPE (SD) 6.7 (5.7) to 74.3 (12.8)% | Activity over 7-days under everyday conditions |
| Reid, 2017 | Fitbit | Flex | Actigraph GT3X+ | Mean (SD) bias (Bland-Altman) -57.5 (46.4) min/day, with mean 64.6 min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Sirard - Phase 3 (Field), 2017 | Movband | Movband | ActiGraph GT3X+ | Spearman's r 0.76 | Activity over 4-days under everyday conditions |

| | | | | | |
|---|---|---|---|---|---|
| Sirard - Phase 3 (Field), 2017 | Sqord | Sqord | ActiGraph GT3X+ | Spearman's r 0.86 | |
| St-Laurent, 2018 | Fitbit | Flex | Actigraph GT3x | Mean (SD) bias (Bland-Altman) 2.4 ± 6.6 (p = 0.21) min/day, with mean (SD) 9.9 (7.5) min/day measured with the reference standard | Activity over 7-days under everyday conditions |
| Alharbi, 2016 | Fitbit | Flex | Actigraph | Mean percentage error 10% | Activity over 4-days under everyday conditions |
| Imboden, 2018 | Fitbit | Flex | ActiGraph GT3X+ | Mean percentage error -65% | Participated in an 80-minute protocol of exercises in a laboratory condition |
| Imboden, 2018 | Jawbone | Up24 | ActiGraph GT3X+ | Mean percentage error -35% | |
| Outcome: Physcial activity intensity | | | | | |
| Bulathsinghala, 2014 | ActiGraph | GT3X+ | ActiGraph GT3X+ on the waist | Physical activity intensity above the threshold was present in16% of the recorded minutes. Mean Vector Magnitude Unit (VMU - movement in three planes) from the wrist device above the 3000 threshold were 4953 (95% confidence interval (CI), 4850 to 5055), while corresponding VMU from the waist device were 951 (95% CI, 916 to 986). Using a proprietary software equation developed for the waist location, activity intensity above this threshold corresponded to 1.66 metabolic units (METs) (95% CI, 1.55 to 1.77). | Activity over 24 hours under everyday conditions |
| Outcome: Speed | | | | | |
| Cohen, 2010 | ActiGraph | Mini MotionLogger | Actual speed (distance/time) | Mean difference 0.97 mph (95% CI, 0.73 - 2.67) | Completed a standardized sequence of activities that comprised sitting, standing, and walking in laboratory setting |

| | | | | | |
|---|---|---|---|---|---|
| Outcome: Step count | | | | | |
| Stackpool, 2013 | Jawbone | UP | Manual count | Pearson's r 0.34 to 0.99 | First session completed on a treadmill at walking or running speed, selected by the participant. Second session was on elliptical cross-trainer at self-selected speed. Apart of the second session also took place in a gymnasium, where they completed agility ladder drills, basketball throws, and basketball lay-ups. |
| Stackpool, 2013 | Nike | Fuelband | Manual count | Pearson's r 0.17 to 0.98 | |
| Stackpool, 2013 | Fitbit | Ultra | Manual count | Pearson's r 0.44 to 0.99 | |
| An, 2017 | Fitbit | Flex | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 4.7 to 21.9% in lab, 18.1% on field | Walking/jogging on a treadmill, walking over-ground on an indoor track, and a 24-hour free-living condition |
| An, 2017 | Garmin | Vivofit | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 2.4 to 16.5% in lab, 17.8% on field | |
| An, 2017 | Polar | Loop | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 9.9 to 23.8% in lab, 26.9% on field | |
| An, 2017 | Basis | B1 Band | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 3.1 to 9.0% in lab, 18.4% on field | |
| An, 2017 | Misfit | Shine | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 6.3 to 19.3% in lab, 23.3% on field | |
| An, 2017 | Jawbone | UP24 | Manual count (Tally Counter) for lab setting, New Lifestyle | MAPE 2.9 to 7.0% in lab, 27.9% on field | |

| | | | | | |
|---|---|---|---|---|---|
| | | | (NL-1000 Series) pedometer for field setting | | |
| An, 2017 | Nike | FuelBand SE | Manual count (Tally Counter) for lab setting, New Lifestyle (NL-1000 Series) pedometer for field setting | MAPE 10.2 to 45.0% in lab, 16.0% on field | |
| Gaz, 2018 | Fitbit | Charge HR | Manual count (Tally counter) | Mean (SD) difference 21.81 (67.08) to 195.06 (207.94) steps. Max distance 1.6 km. | Performance on a free walking or treadmill walking condition. Treadmill walking had pre-determined speeds |
| Gaz, 2018 | Apple | Watch, series not NA | Manual count (Tally counter) | Mean (SD) difference 7.56 (29.61) to 39.44 (151.81) steps. Max distance 1.6 km. | |
| Gaz, 2018 | Garmin | Vivofit 2 | Manual count (Tally counter) | Mean (SD) difference 5.09 (8.38) to 98.06 (137.49) steps. Max distance 1.6 km. | |
| Gaz, 2018 | Jawbone | UP2 | Manual count (Tally counter) | Mean (SD) difference 16.19 (29.14) to 64 (66.32) steps. Max distance 1.6 km. | |
| Hargens, 2017 | Fitbit | Charge | ActiGraph GT3x | MAPE 20.7% | Activity over 7-days under everyday conditions |
| Jones, 2018 | Fitbit | Flex | Manual count (video) | MAPE 0-4% | Completed treadmill protocol at jogging and running speeds (8km/h-16km/h) in laboratory settings |
| Lauritzen, 2013 | Fitbit | Ultra | Manual count (video) | MAPE (SD) 99.6 (0.8)% | Walking procedure of a straight path over 20m in a laboratory setting |
| Magistro, 2018 | ADAMO | Care Watch | Manual count (Tally counter) | MAPE (SD) -17.70 (20.77) % to -1.10 (2.30) % | Performance on randomly ordered tasks: walking slow, normal and fast self-paced speeds, and up/down stairs in a laboratory setting |
| Montoye, 2017 | Fitbit | Charge HR | Omron HJ 323u Pedometer (Omron Corp., Osaka, Japan) | MAPE (SD) 9.7% (1.2) | Performing 14 activities in a laboratory and on a track (lying, sitting, standing, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | walking various speed and inclines, jogging, and cycling) |
| Alsubheen, 2016 | Garmin | Vivofit | Kinematics analysis (video camera Sony-HDR-FX1 12X HD, Mini DV Camcorder) | Vivofit systematically underestimated step count only at 0% treadmill inclination | Performance on treadmill walking tasks, and office activities within a laboratory session, completed in separate sessions on different days |
| Falgoust, 2018 | Fitbit | Charge HR | Manual count (Tally counter) | Mean difference - 60.8 steps (p 0.01) | Performance on track laps in laboratory settings |
| Falgoust, 2018 | Fitbit | Surge | Manual count (Tally counter) | Mean difference -86.0 steps (p 0.004) | |
| Falgoust, 2018 | Garmin | Vivoactive HR | Manual count (Tally counter) | Mean difference -19.7 steps (p 0.03) | |
| Blondeel, 2018 | Fitbit | Alta | Dynaport Movemonitor (accellerometer) | Mean difference (SD) 773 (829) steps (p=0.009) | Activity over 14-days under everyday conditions |
| Boeselt, 2016 | Polar | A300 | BodyMedia SenseWear | Pearson's r 0.96 ($p < 0.01$) | Performance in everyday conditions |
| Burton, 2018 | Fitbit | Flex | Manual count (video) by two researchers | Intraclass Correlation (ICC) 0.77 (0.57,0.88) and 0.76 (0.53,0.88) in two 2-minutes walking tests | Two 2-minute walk tests were completed while wearing the fitness trackers. Participants were videoed during each test. Participants were then given one fitness tracker and an accelerometer to wear at home for 14-days. |

# Supplementary material – Chapter 4

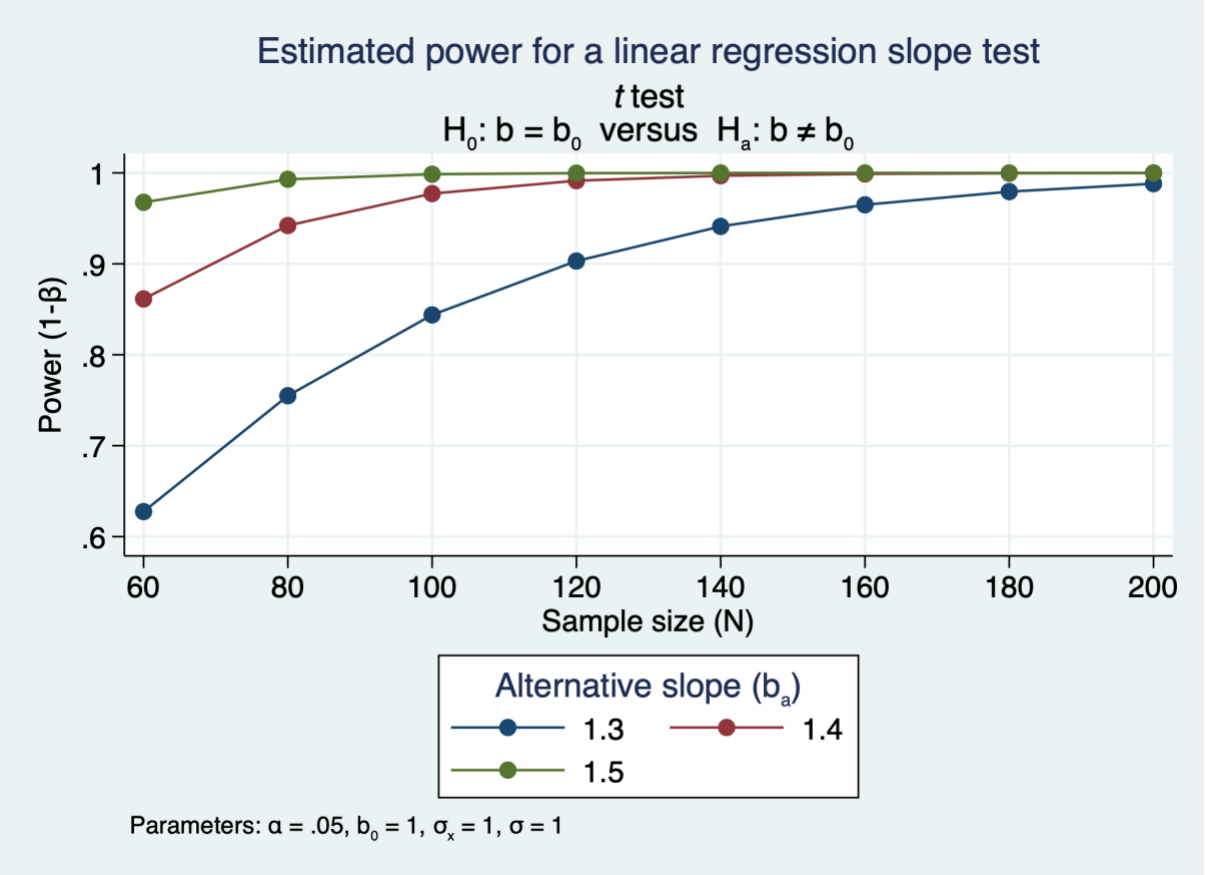
## Methods

*sTable 17: mapping between factor concentrate and Population PK model used.*

| Factor | Class | Concentrate | Model |
|---|---|---|---|
| FVIII | SHL | Advate | Generic for FVIII Full Length Recombinant |
| FVIII | SHL | Humate-P | Generic for FVIII Plasma Derived |
| FVIII | SHL | Kovaltry | Generic for FVIII Full Length Recombinant |
| FVIII | SHL | Nuwiq | Generic for FVIII B Domain Deleted |
| FVIII | SHL | Recombinate | Generic for FVIII Full Length Recombinant |
| FVIII | SHL | Wilate | Generic for FVIII Plasma Derived |
| FVIII | SHL | Xyntha | Generic for FVIII B Domain Deleted |
| FVIII | SHL | Zonovate | Generic for FVIII B Domain Deleted |
| FVIII | EHL | Adynovate | Adynovate |
| FVIII | EHL | Eloctate | Elocta/Eloctate |
| FVIII | EHL | Jivi | Jivi |
| FIX | SHL | BeneFIX | Generic for FIX Recombinant |
| FIX | SHL | Immunine VH | Generic for FIX Plasma Derived |
| FIX | SHL | Rixubis | Generic for FIX Recombinant |
| FIX | EHL | Alprolix | Alprolix |
| FIX | EHL | Idelvion | Idelvion |
| FIX | EHL | Rebinyn | Rebinyn |

FVIII: factor VIII; FIX: Factor IX, SHL: standard half-life; EHL: extended half-life.

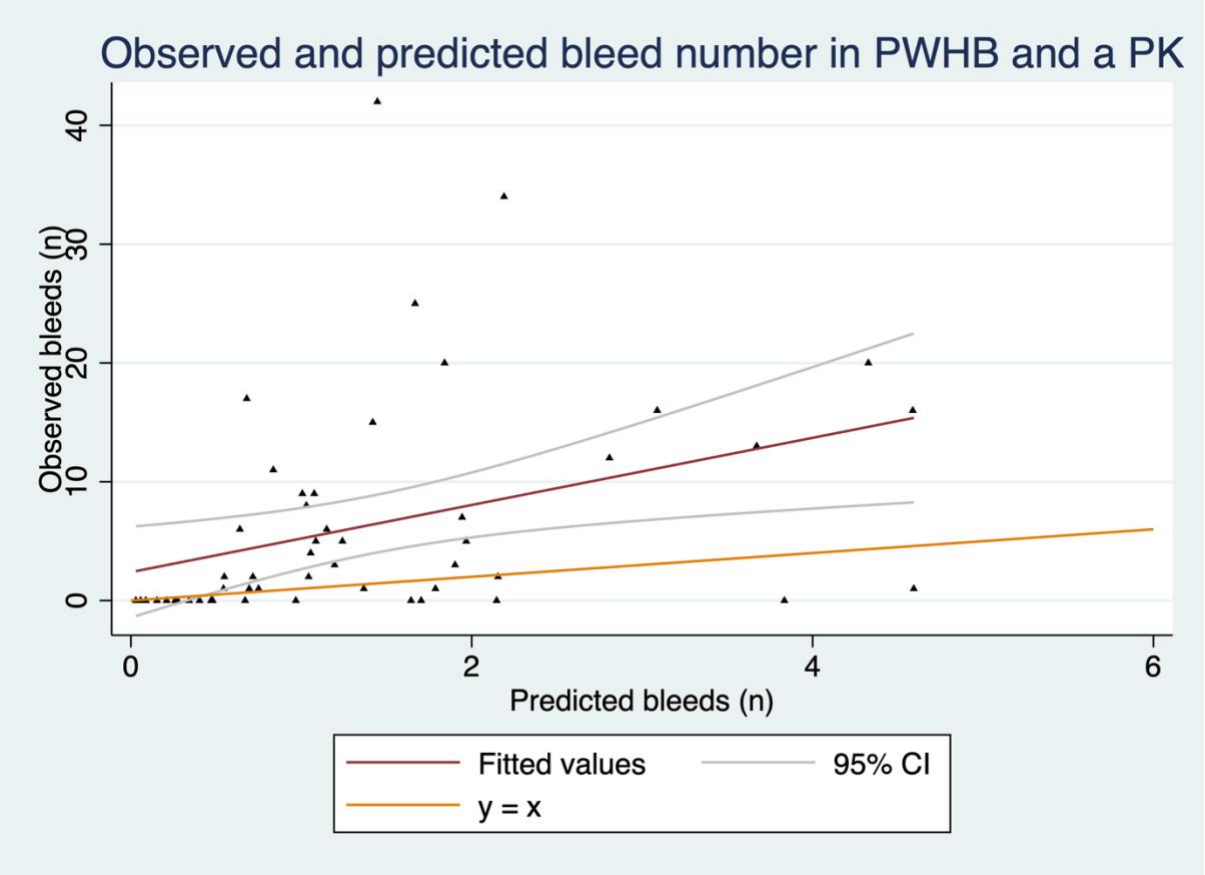
*sFigure 10: estimated power for different scenarios.*




## Results: model performance

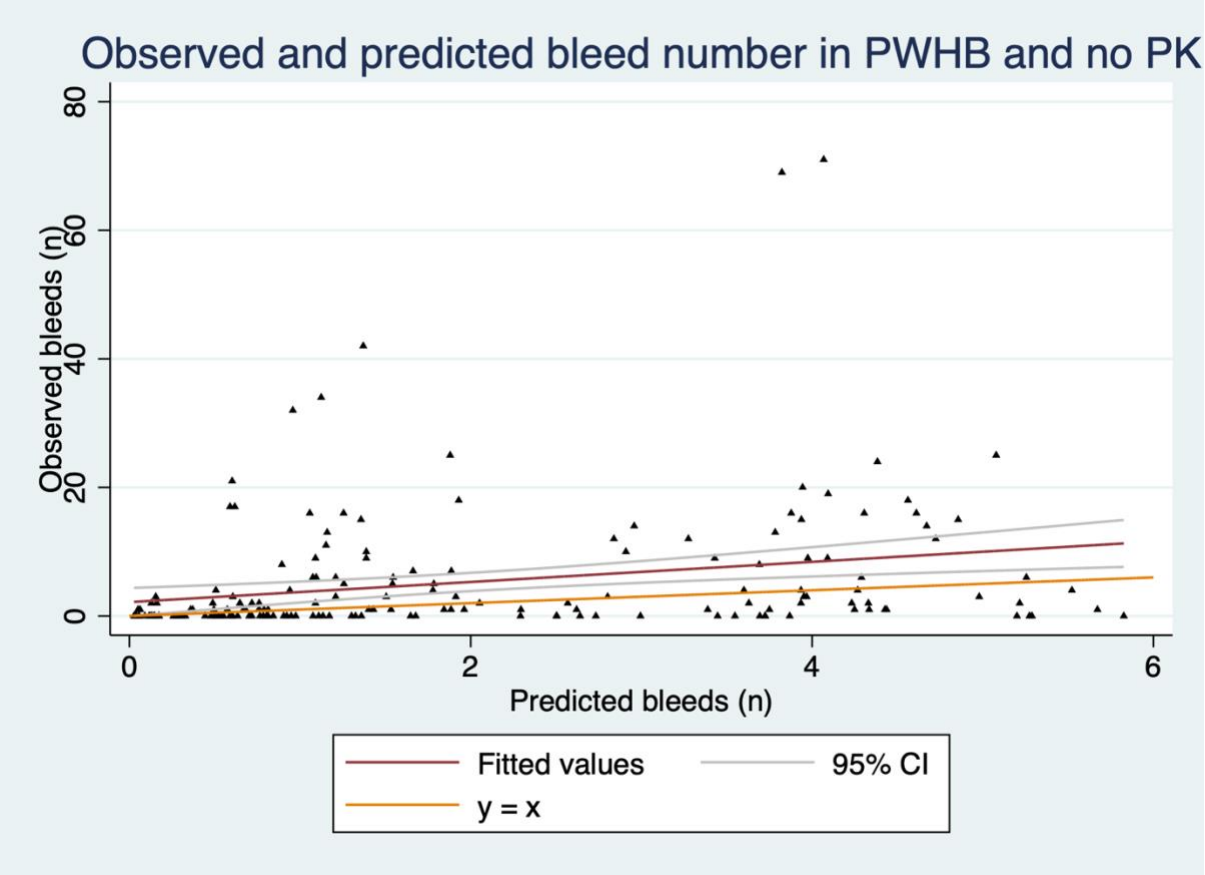
### Predicting the total number of bleeds

*sFigure 11: observed versus predicted number of bleeds during each treatment period, for PWH B with an individual PK available.*




CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

*sFigure 12: observed versus predicted number of bleeds during each treatment period, for PWH B without an individual PK available.*




CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

*sTable 18: percentage of the observations falling within the 5th and 95th percentiles of the prediction, for the total number of bleeds.*

| | n | Percentage | 95% CI |
|---|---|---|---|
| PWH A with PK | 125/151 | 82.8% | 76.8%; 88.8% |
| PWH B with PK | 30/51 | 58.8% | 45.3%; 72.3% |
| PWH A no PK | 446/632 | 70.6% | 67.0%; 74.1% |
| PWH B no PK | 112/180 | 62.2% | 55.1%; 69.3% |

CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

**Observed and predicted Kaplan-Mayer curves for the occurrence of the first bleed**

*sFigure 13: observed versus predicted survival time (first bleed) during each treatment period, for PWH B and an individual PK available.*

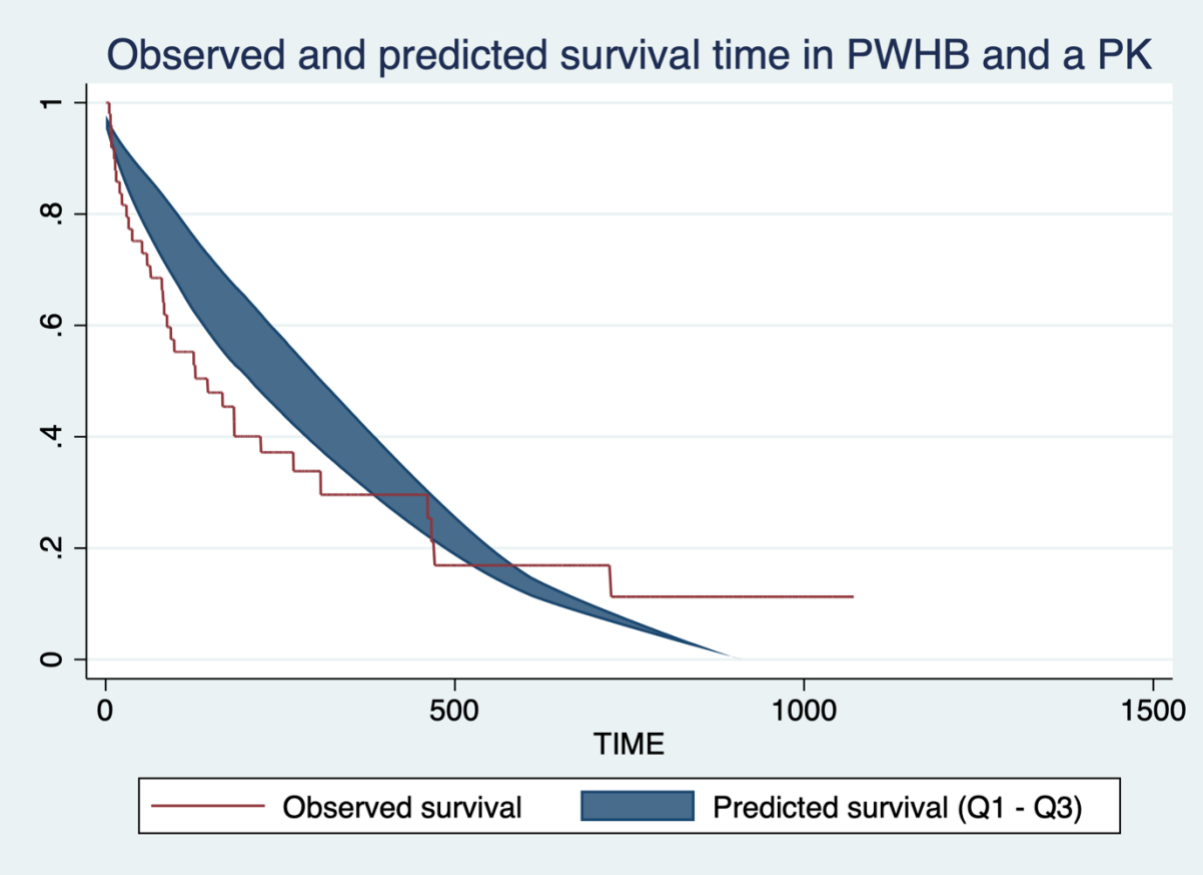
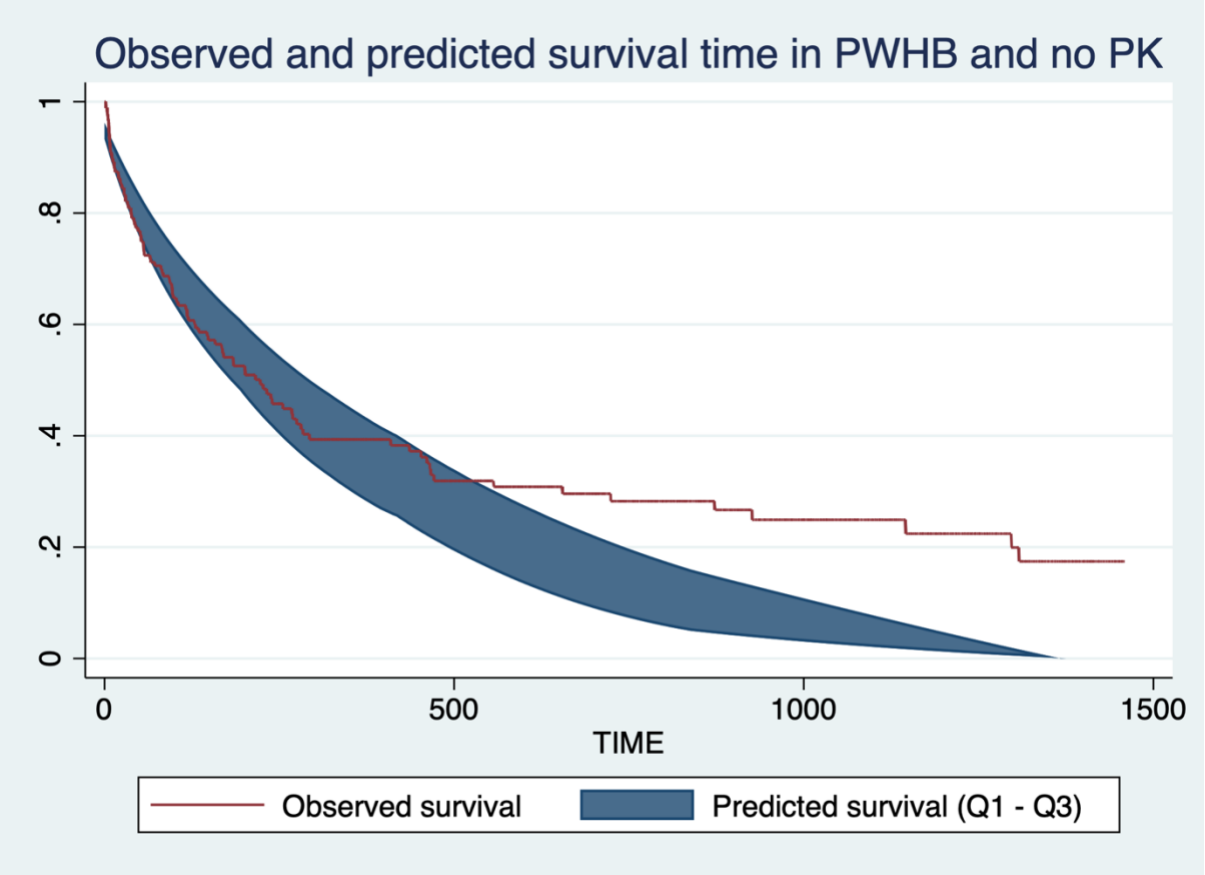



*Figure 14: observed versus predicted survival time (first bleed) during each treatment period, for PWH B without an individual PK available.*



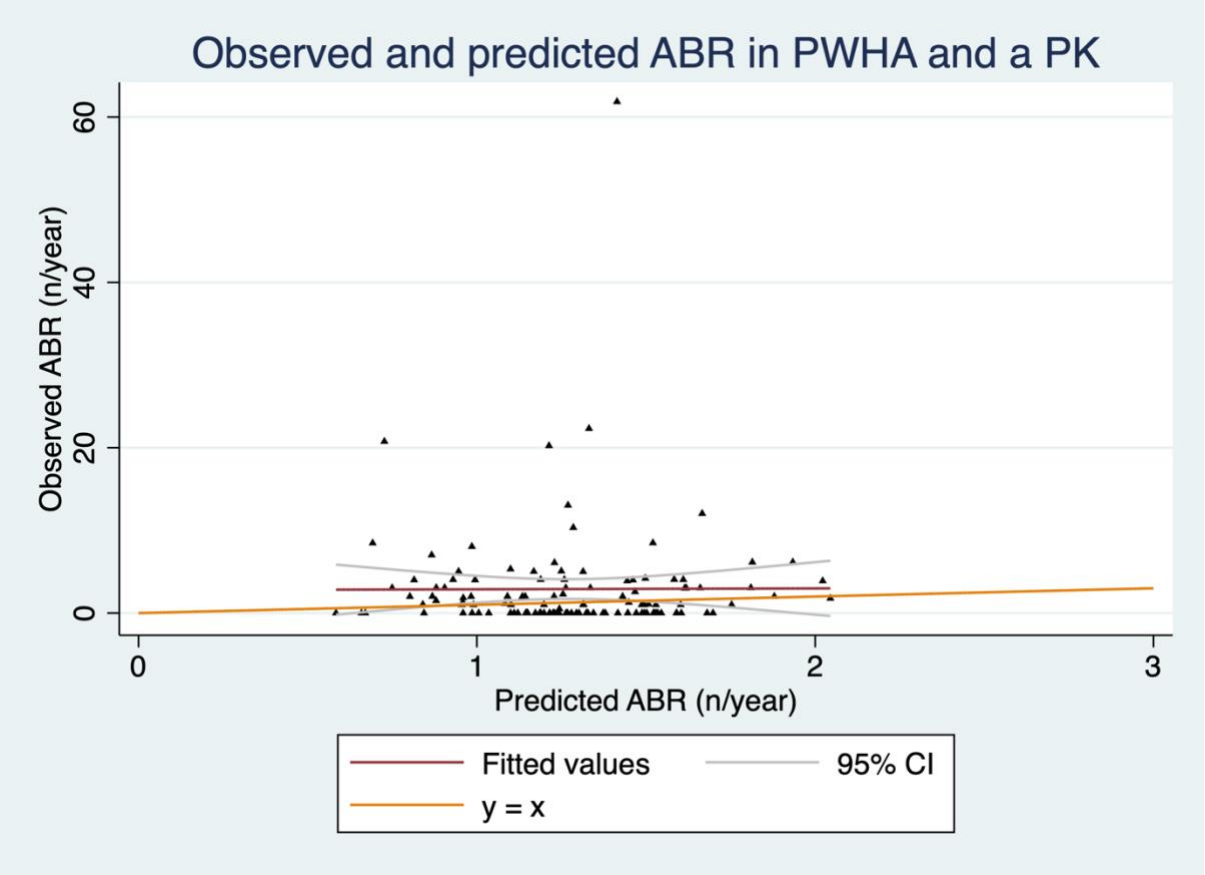

**Predicting the annualized bleeding rate (ABR)**

*sTable 19: regression analysis for observed versus predicted ABR.*

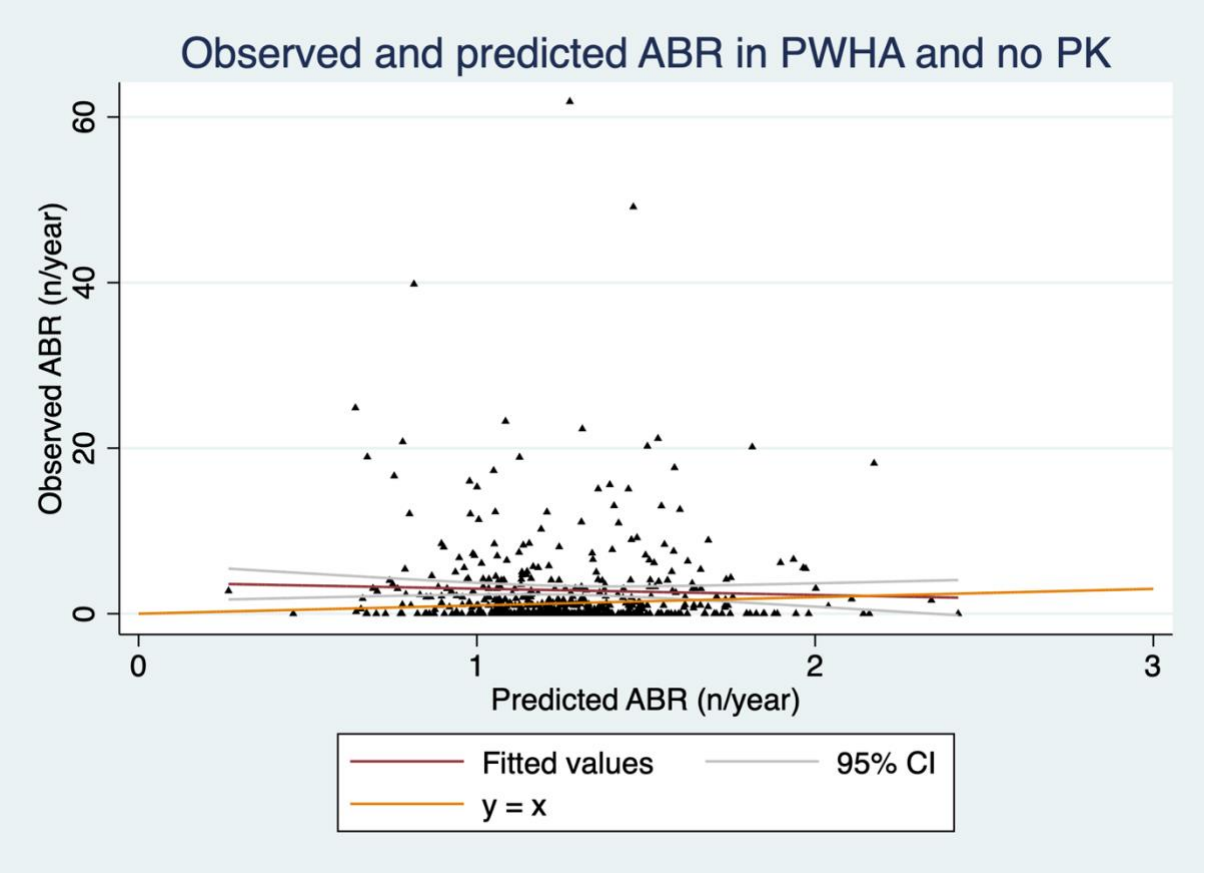
| | PWH A with PK | PWH B with PK | PWH A no PK | PWH B no PK |
|---|---|---|---|---|
| n | 125 | 41 | 459 | 132 |
| Coefficient (95% CI) | 0.09 (-2.75; 2.93) | 1.26 (-2.44; 4.95) | -0.76 (-2.90; 1.38) | 0.06 (-1.35; 1.47) |
| p value | 0.950 | 0.492 | 0.485 | 0.931 |
| Intercept (95% CI) | 2.76 (-0.70; 6.21) | 3.10 (0.14; 6.04) | 3.78 (0.89; 6.68) | 3.40 (1.40; 5.40) |
| p value | 0.117 | 0.040 | 0.010 | 0.001 |
| R2 | 0.00 | 0.01 | 0.00 | 0.00 |

CI: confidence interval; PK: pharmacokinetics; PWH: people with hemophilia.

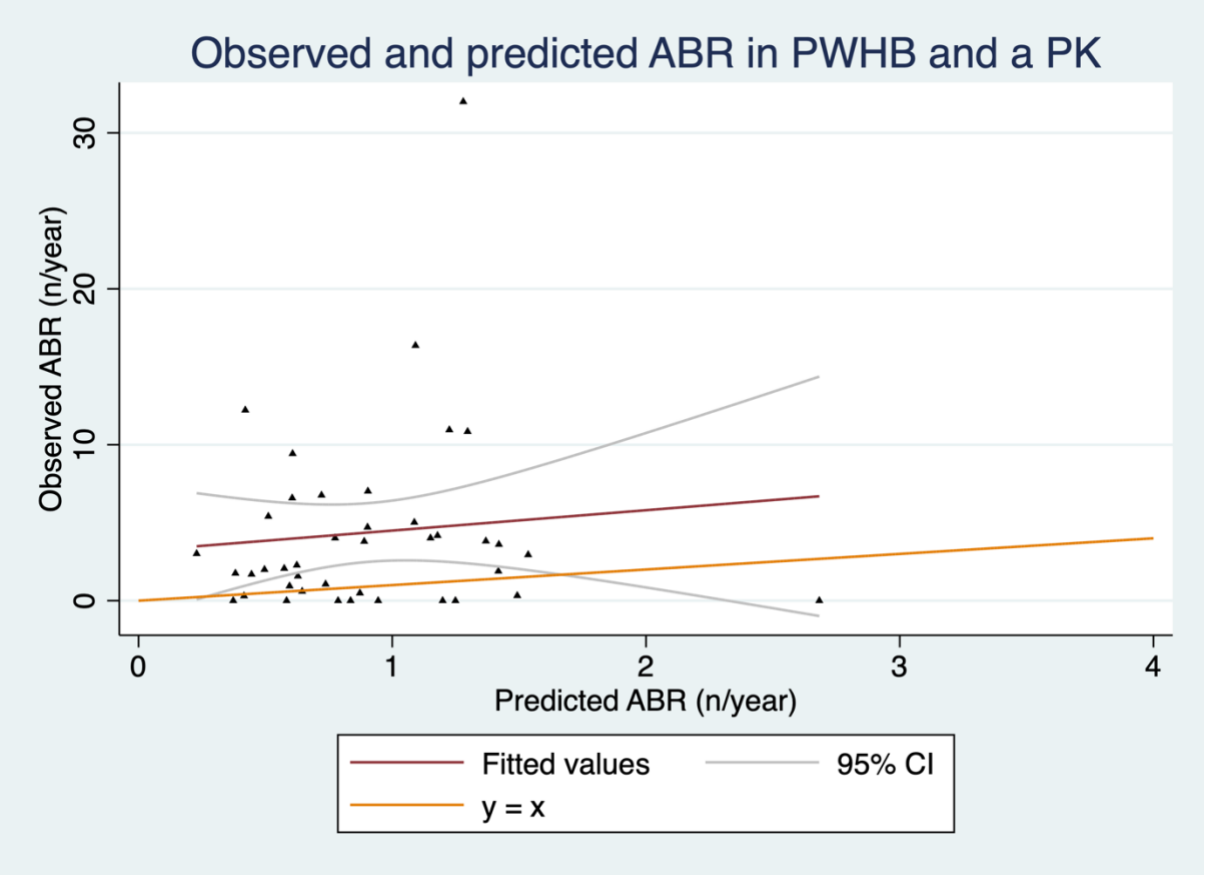
*sFigure 15: observed versus predicted ABR during each treatment period, for PWH A with an individual PK available.*



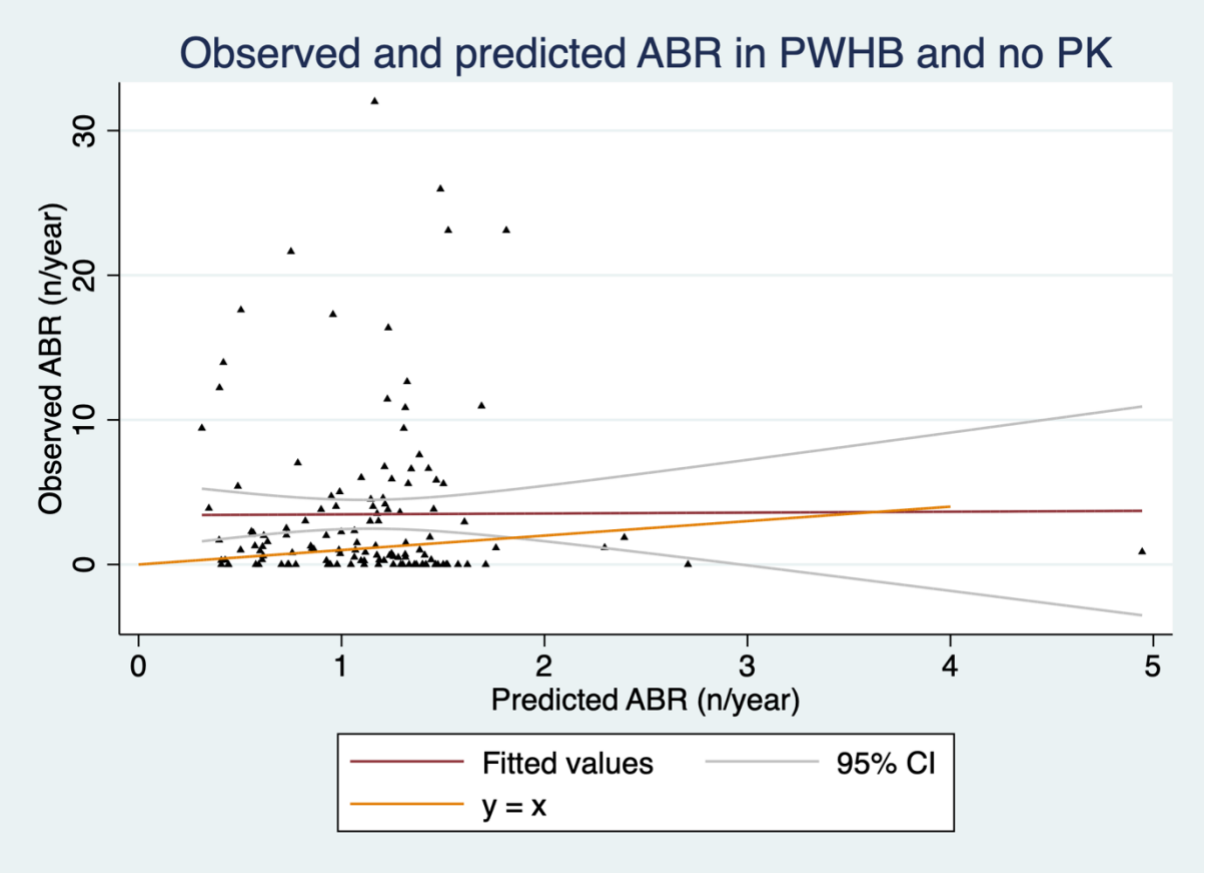

*sFigure 16: observed versus predicted ABR during each treatment period, for PWH A without an individual PK available.*




*sFigure 17: observed versus predicted ABR during each treatment period, for PWH B with an individual PK available.*




*sFigure 18: observed versus predicted ABR during each treatment period, for PWH B without an individual PK available.*




*sTable 20: accuracy for predicting an ABR ≥ 4 bleeds per year.*

| | PWHA with PK | PWHB with PK | PWHA no PK | PWHB no PK |
|---|---|---|---|---|
| N | 125 | 41 | 632 | 180 |
| Sensitivity | NA | NA | NA | 0.0% |
| Specificity | 78.4% | 63.4% | 79.7% | 74.1% |

NA: not available; PK: pharmacokinetics; PWH: people with hemophilia.

# Appendix to the supplementary material – Chapter 4 – derivation of the prediction model

**Title**: Estimated Factor VIII Activity Levels at the Time of Bleeding Events in Individuals with Hemophilia A Without Inhibitors.

Short title:

**Authors**: Pierre Chelle[1], Dagmar Hajducek[1], Nabil Daoud[2], Emma Iserman[3], Christine Gerber[2], Pratik Bhagunde[4], Suresh Katragadda[4], Federico Germini[3], Alfonso Iorio[3], Andrea Edginton[1], Michael Recht[2]

**Affiliations**:

[1] School of Pharmacy, University of Waterloo, Waterloo, Ontario, Canada;

[2] American Thrombosis and Hemostasis Network, Inc., Rochester, New York, United States of America;

[3] Dept of Health Research Methods, Evidence and Impact and Dept of Medicine, McMaster University, Ontario, Canada;

[4] Sanofi, Bridgewater, NJ, United States of America.

Target journal: Blood

**Type of article**: Regular Article / **Scientific Category**: Thrombosis & Hemostasis

Abstract word count: <250 words

Text word count: <4000 words

Table/Figure count: <= 7 figures

Corresponding Author: Andrea Edginton

## Abstract

**Background:** People with hemophilia A (PwHA) lack clotting factor VIII (FVIII) resulting in bleeding events (BEs) which may be prevented by maintaining a factor level above a target threshold using pharmacokinetic (PK)-driven prophylaxis.

**Aims:** Assess the relationship between FVIII levels and bleeding risk in PwHA using real world data.

**Methods:** Real world data was retrospectively collected from the American Thrombosis and Hemostasis Network dataset, the Web Accessible Population Pharmacokinetics Service – Hemophilia (WAPPS-Hemo) database and the Canadian Bleeding Disorders Registry, limited to participants with PK, BEs and dose information. BEs were classified according to cause (trauma or spontaneous) and location (joint or non-joint). PK parameters were obtained by Bayesian estimation using WAPPS-Hemo population PK models. FVIII levels were simulated using PwHA's dose records and estimated PK parameters. FVIII – bleeding risk relationship was assessed by developing and evaluating a repeated time to event (RTTE) model.

**Results:** 2862 BEs from 427 PwHA aged 1 to 66 years with median of 21 years were identified; including 1434 BEs related to trauma, 1428 spontaneous and 1984 joint bleeding events. Observation period ranged from 3 weeks to 11 years with median of 2.28 years (Table 1). Simulated FVIII was below 1 IU/dL for 864 (30.2%) bleeding events and between 1 and 10 IU/dL for 1295 (45.2%) bleeding events (Figure 1). The RTTE model included Hill and a Weibull models

in the hazard function describing the effects of FVIII levels and observation time on bleeding risk, and between-subject variability (BSV) on base hazard (CV=79%), bleed cause probability (CV=59%) and bleed location probability (CV=100%). Agreement between observed vs estimated bleed count was $R^2$=0.95.

**Conclusion:** Trough levels were well correlated with bleeding risk independent from dosing.

## Introduction

People with hemophilia A are deficient in or lacking clotting factor VIII, a necessary protein for the body to create stable blood clots at the site of injury, resulting in bleeding episodes occurring spontaneously or as a result of trauma. Without prophylaxis with either factor VIII replacement or factor VIII mimetics, people with hemophilia can be expected to experience over 20 bleeding episodes a year [1]. Over the past 20 years, it has become clear that prevention of bleeding is best accomplished with either scheduled administration of intravenous factor VIII concentrate or subcutaneous administration of a factor VIII mimetic [2,3]. However, these medications do not completely abrogate the occurrence of "breakthrough" bleeding episodes. Recently, administration of pharmacokinetic-based factor VIII concentrate prophylaxis has become more common. While in the past, the goal of prophylaxis was to keep the factor VIII activity level above 1%, pharmacokinetic-based factor VIII concentrate prophylaxis, particularly when using extended half-life factor VIII concentrates aims for higher trough levels, in the 10-15% range [4]. While theoretically appealing, there is as yet no trials demonstrating the superiority of maintaining a trough factor VIII activity in this higher range.

This study is a collaborative work between the American Thrombosis and Hemostasis Network (ATHN), a patient health network gathering data from individuals with bleeding and clotting disorders receiving care through hemophilia treatment centers across the United States, and the Web Accessible Population Pharmacokinetics Service – Hemophilia (WAPPS-Hemo), a web platform applying population pharmacokinetics (PopPK) to the personalisation of prophylactic

therapy for the treatment of hemophilia, promoting patient awareness and involvement in the management of their illness. The study was performed using data from ATHN dataset, WAPPS-Hemo database and the Canadian Bleeding Disorders Registry (CBDR), a clinical database for patients in Canada with bleeding disorders aimed at assisting in managing the treatment of people with bleeding disorders.

The objective of the work was to assess the relationship between factor VIII activity levels and bleeding risk and further to determine if and how specific covariates including treatment, bleed cause and bleed location affect bleeding risk.

# Methods

## Patients and data

### Sources

Data required for PK assessment was extracted from the ATHN dataset and the WAPPS-Hemo database in April 2021. These data included mandatory population PK model covariates, namely body weight, height, age and baseline endogenous factor level; infused factor VIII concentrate, amounts and times of doses, and factor VIII measurements and times of PK samples.

Bleed data matching the same participants and concentrates of the PK dataset was also extracted from the ATHNdataset database and the CBDR database in April 2021. Bleed data included recorded amounts and times of concentrate doses, as well as recorded times of bleeding events, bleed causes (ie,. traumatic, spontaneous) and bleed location (i.e. joint, non-joint). Observation periods corresponded to the time frame in which records of doses and bleeds were available.

Due to changes in PK parameters and treatment plans when patients switch to a different factor concentrate class, every patient that switched to a different factor concentrate class was assumed as a different participant in the analysis.

**Inclusion/Exclusion criteria**

All hemophilia A participants for whom a PK study had been completed on a concentrate for which bleeds had been tracked were eligible for inclusion in the study.

Participants with any bleeding disorder in addition to or other than hemophilia A were excluded from the study. To ensure that the Bayesian estimation of PK led to reliable PK parameters, PK studies that included only one PK sample within the first four hours after dose were not included. Administration of any non-FVIII concentrate to manage prophylaxis may lead to unreliable PK, thus bleeds data in that time frame were removed from the study. Likewise, bleeds data within the time frame of positive inhibitors status were excluded.

## PK analysis

The population pharmacokinetic modelling and the Bayesian estimation methodology were executed using NONMEM and PDx-Pop (ICON Development Systems, Ellicott City, MD, USA; v 7.4.0 and v 5.10, respectively). Statistical and graphical analyses of the models were performed using R (R Project, v 3.6.1).

PK parameters for each patient were obtained from the PK study by means of Bayesian forecasting using the population PK models developed in WAPPS-Hemo [5]. Since most of these population PK models are concentrate-specific, the mapping between models and concentrates is indicated below:

- For standard half-life (SHL) concentrates
    - Generic FVIII Recombinant model was used for recombinant concentrates.

  - Generic FVIII Plasma Derived model was used for plasma derived concentrates.
  - Generic FVIII Recombinant BDD model was used for recombinant B domain-deleted concentrates.
- For extended half-life (EHL) concentrates
  - Adynovate model was used for the EHL concentrate Adynovate.
  - Eloctate model was used for the EHL concentrate Eloctate.
  - Jivi model was used for the EHL concentrate Jivi.

All the population PK models were 2-compartment PK models that included between subject variability on clearance (CL) and central volume (V1). Age was a covariate on CL and fat-free mass a covariate on CL and V1.

Estimated PK were evaluated by assessing the agreement between observed and simulated FVIII activities as performed in Simulated vs Observed plots and prediction-corrected visual predictive check plots [6].

## Bleed analysis

### Analysis of FVIII at time of bleeding event

For each patient, FVIII activity over time was simulated from the PK parameters previously estimated, along with dosing information recorded in the bleed data. To assess FVIII activity at the time of bleeding events, distributions of the relevant FVIII activities were graphically displayed using histograms split and compared according to covariates of interest: concentrate class (SHL vs EHL), age (<12 yrs vs ≥12 yrs), body mass index (<30kg/$m^2$ vs ≥30kg/$m^2$), bleed

category (traumatic vs spontaneous), and bleed location (joint vs non-joint). Statistical comparison of the distribution was assessed using a Wilcoxon test; accounting for the sample size, effect size and distribution of the data, a p-value of 0.01 was assumed significant [7].

**Analysis of survival**

Survival is defined as the probability that a participant is event-free beyond a specific time and can be expressed in terms of a hazard function, defined as the rate in which a participant experiences an event by a specific time. In the context of repeated bleeding events, the survival was assessed by the Kaplan-Meier estimator, corresponding to the fraction of participants that have had a bleed (or a number of bleeds) at a specific time with respect to the number of participants still at risk of experiencing a bleed (or a number of bleeds) at the same specified time. Participants still below a bleed count at the end of their observation period were removed from the population still at risk of experiencing bleeding events. This censoring of the data was graphically displayed as vertical ticks in the plots of the Kaplan-Meier estimator. Plots of the Kaplan-Meier estimator for various combinations of covariate categories were performed. The graphical analysis was done using functions from the R package *vpc* [8].

**Repeated time to event population modeling**

Repeated time to event (RTTE) population modeling aimed at describing survival – i.e. time until bleeding events occur - and its variability within a population. This was performed by first defining a hazard function (*h*) describing a rate for bleeding events to occur. Hazard was then converted into a survivor function and the likelihood necessary for parameter estimation. The

modeling corresponds to the investigation of the relationship between hazard, its variability and factors influencing the times until bleeding events occur.

A detailed description of the model development and its evaluations are presented in Supplemental Material. The main steps of these methods are summarized below.

Model development and evaluations

RTTE model development consisted of the selection of the best structural model followed by a covariate analysis. Observation time and simulated FVIII activity were shown to affect hazard in previous RTTE models [9–11]. The structural model selection corresponded to the inclusion of the models best describing the relationships between (i) observation time and hazard and (ii) simulated FVIII activity and hazard.

To describe the relationship between observation time and hazard, exponential, Gompertz and Weibull models were tested. To describe the relationship between simulated FVIII activity and hazard, constant (no effect), Hill and Power models were tested (the equations for each model are presented in Supplemental Data).

Covariate analysis corresponded to the assessment of the effect of Age, BMI, and concentrate class (SHL vs. EHL) on every model parameter that included between subject variability (BSV). Covariate effects were normalized by the median covariate value and modelled based on the shape of the scatter plot describing the covariate vs BSV parameter of interest (e.g. linear, power or log effect). Forward inclusion was assumed significant if the objective function dropped at least by 3.84 ($p<0.05$), and backward elimination was assumed significant if the objective function increased at least by 6.64 ($p<0.01$)

Evaluations of the population model were based on its scientific plausibility – checking that the estimated values of model parameters were in expected range for hazard and effects of time and factor VIII activity levels, changes in objective function values (OFV) and information criteria (OFV: -2log-likelihood, AIC and BIC), goodness of fit plots and precision of parameter estimates; in accordance with the tutorial [12] and a previous published RTTE model of hemophilia bleeds [9]. BSV was evaluated through the assessment of $\eta$ distributions and their shrinkage values. Since the model objective was descriptive rather than predictive, values of $\eta$-shrinkage of up to 50% were considered satisfactory. However, no interpretation of $\eta$ vs. covariates plots was proposed when $\eta$-shrinkage was higher than 20% [13]. Model evaluations also included dedicated methods for RTTE modeling such as Visual Predictive Checks (VPCs) of the Kaplan-Meier estimator, stratified by occurrence of first, second and third bleed, and stratified by bleed cause/location for the first bleeding event (detailed explanations of the VPC simulations are presented in Supplemental Data).

Simulations

Simulations were performed using the final population RTTE model to highlight the effect of treatment regimen such as dose and frequency of infusions on bleeding risk. The simulations used generic virtual patients with demographic and PK parameters defined from the study population. The population of real participants was split by concentrate class and age to simulate 4 generic virtual patients obtained by calculating median demographic and PK parameters. Virtual patient (1) corresponded to participants ≥ 12 yrs infused with SHL concentrates, virtual patient (2) corresponded to participants < 12 yrs infused with SHL

concentrates, virtual patient (3) corresponded to participants ≥ 12 yrs infused with EHL concentrates, and virtual patient (4) corresponded to participants< 12 yrs infused with EHL concentrates. A typical value of hazard ($\eta$=0) was used to simulate the median relationship between factor activity and bleeding risk. Bleeding risk variability in the population was simulated using hazard BSV ($\eta$=±1.645*standard deviation) defining a population range delimited by the $5^{th}$ and $95^{th}$ percentiles of hazard BSV.

A first simulation exercise derived the hazard and survival for generic patient (1) infused 35 IU/kg every 3 days for 2 years. The outcome was compared to that of a literature model (Abrantes *et al.* [9]) using the same generic patient and dosing regimen.

A second simulation exercise compared the effect on bleed risk of 5 different prophylactic treatments simulated for 2 years on each generic patient. The following PK and PD outcomes were reported for each simulation: the minimum and average factor activities (Cmin and Cavg), the time spent above 3 IU/dL (TAT3), the hazard mean value, the probability of having at least one bleed within 0.5, 1 and 2 years for typical and $5^{th}$-$95^{th}$ percentile population range of hazard BSV. Simulated treatments were:

a) Reference dose: 35 IU/kg every 3 days
b) Higher dose: 50 IU/kg every 3 days
c) Lower dose: 20 IU/kg every 3 days
d) Higher frequency: 35 IU/kg every 2 days
e) Lower frequency: 35 IU/kg every 4 days

# Results

## Patients and data

The final dataset included 436 PK studies from 427 participants. Some participants had multiple PK studies available for the same concentrate such that the number of PK studies was greater than the number of participants summarized in the bleed dataset.

The data included patients ranging from 9 months to 68 yrs, with a median of 20.5 yrs. Participants < 12 years of age comprised 28.7% of the PK studies (Table 1). Overall, 16.1% of the PK studies were in participants with mild or moderate hemophilia. A history of, but not active, inhibitors was present in 12.4% of participants and more prevalent in children <12 years. Sixty-one percent of PK data were related to an SHL concentrate. The age distribution was similar between SHL and EHL PK studies.

In the PK dataset, a median of 3 PK samples were available per patient with no trend between age or concentrate class. Weight-normalized dose was higher for EHL concentrates compared to SHL concentrates, on average (Table 1).

The bleed dataset had a heterogeneous range of observation periods from 3 weeks to 11 years with a median of 2.28 yrs. Participants using EHL concentrates had a shorter observation period than participants using SHL concentrates with median [min-max] of 1.13 [0.04 – 11.22] years against 2.71 [0.07-11.22] years, respectively. The data were also heterogeneous in terms of recorded dosing regimen with a median [min-max] dose of 34.3 [7.0-321.4] IU/kg and median frequency of 2.96 [0.34-395.2] days (Table 2). Participants using EHL concentrates had a median

dose of 37.3 [10.0-123.9] IU/kg and median frequency 3 [1-123] days, while participants using SHL concentrates had a median dose of 32.0 [7.0-321.4] IU/kg and a median frequency 2.1 [0.3-395] days.

In total, 2862 bleeds were recorded, approximately half spontaneous and half trauma related. Regardless of bleed cause, 66.2 % of the total bleeds were at joint locations (Table 3). By participant, the bleed count at the end of the observation period had a median [min-max] of 2 [0-199] bleeds. This was translated in terms of annualized bleeding rate (ABR) to 1.1 [0.0-37.1] bleeds per participant per year. ABR was higher in children compared to adults, however not significantly (median ABR of 1.34 vs 1.04, $p = 0.395$). ABR was significantly higher for participants using SHL concentrates compared to participants using EHL concentrates (median ABR of 1.47 vs 0.41, $p<0.01$). Split by cause and location, ABR showed similar results. Indeed, regarding joint bleeds only, median ABR was 0.67 for participants using SHL against 0.0 for participants using EHL; regarding non-joint bleeds only, median ABR was 0.36 for participants using SHL against 0.0 for participants using EHL; regarding spontaneous bleeds only, median ABR was 0.37 for participants using SHL against 0.0 for participants using EHL; regarding trauma bleeds only, median ABR was 0.7 for participants using SHL against 0.0 for participants using EHL. From all patients, 32.1 % did not record any bleeding event during their observation period. When splitting by concentrate class, 22.9 % of participants using SHL concentrate did not record any bleeding event while this ratio increased to 46.7 % for participants using EHL concentrates. Likewise, 50.1 % of the participants did not record any spontaneous bleed and 42.2 % did not record any joint bleed. When splitting by concentrate class, 40.8 % and 33.6 % of

participants using SHL concentrate did not record any spontaneous and joint bleed respectively while this ratio increased to 64.8 % and 55.8 % for participants using EHL concentrates (Table 2).

## PK analysis

Evaluations of the Bayesian estimation led to a satisfactory fit of the PK data with coefficient of determination $R^2$=94% (Figure S1) and percentiles of simulated data matching the percentiles of the observed data in the pcVPC (Figure S2 and S3).

Estimated PK parameters were summarized in the second part of Table 1 and split by age and concentrate class to highlight the correlation of PK with age and concentrate class (also illustrated in Figure S4). Briefly, half-life and dose normalized time to 1, 3 and 5 IU/dL were higher in adults regardless of concentrate and were higher in EHL concentrates regardless of age.

## Bleed analysis

### Analysis of FVIII at time of bleeding event

Overall, 864 bleeds (30.2 %) occurred when FVIII was simulated to be below 1 IU/dL and 1295 bleeds (45.2 %) occurred when FVIII was simulated to be between 1 and 10 IU/dL (Figure 1). These ratios were relatively similar between spontaneous and trauma bleeds (Figure S2). From trauma, 451 bleeds (31.5 %) happened when FVIII was below 1 IU/dL and 652 bleeds (45.5 %) happened when FVIII was between 1 and 10%. While for spontaneous, 413 bleeds (28.9 %)

happened when FVIII was below 1 IU/dL and 643 bleeds (45.0 %) happened when FVIII was between 1 and 10 IU/dL.

Distributions of simulated FVIII activity at times of bleeding events split by concentrate class and bleed cause are presented in Figure 2. The overall distribution and the distribution split by cause and location are included in Supplemental Data (Figures S5 and S6). The distributions were not significantly different between trauma and spontaneous bleeds, with median simulated FVIII being 2.8 IU/dL from the 1434 trauma bleeding events and 2.4 IU/dL from the 1428 spontaneous bleeding events, respectively (Wilcoxon $p = 0.04$). The distributions were significant between joint and non-joint bleeds with median simulated FVIII being 2.9 IU/dL from the 1894 joint bleeding events and 1.8 IU/dL from the 968 non-joint bleeding events, respectively (Wilcoxon $p<0.01$). The distributions were also significantly different between bleeds from SHL and EHL usage with median simulated FVIII being 2.2 IU/dL from 2324 bleeding events recorded by the 262 participants using SHL concentrates and 4.4 IU/dL from 538 bleeding events recorded by the 165 participants using EHL concentrates respectively (Wilcoxon $p<0.01$).

The difference in distributions of simulated FVIII at the time of bleeding event between SHL and EHL concentrates was also correlated with a significant difference in average simulated factor VIII activity (Cavg) calculated over the observation period (Figures S7and S8, Wilcoxon $p<0.01$). Patients infused with EHL concentrates spent significantly less time below 5, 3, or 1 IU/dL compared to patients infused with SHL (all Wilcoxon $p<0.01$).

**Analysis of survival**

Survival plots using the Kaplan Meier estimator supported the previous results from the analysis of simulated FVIII activities at time of bleeding events. Briefly, survival curves were similar when split between bleed causes or between bleed locations (Figures 3, and S9 to S13).

**Repeated time to event population modeling**

Model overview

The best structural model described hazard using a Hill function for the effect of FVIII activity and a Weibull function with a negative exponent for time – supporting that more bleeding events were recorded at the early stages of the observation period.

After inclusion of effects of FVIII activity and time on hazard, no covariate was found significant in reducing further the BSV of base hazard (i.e. bleeds/yr). Covariate analysis on bleed cause showed a significant correlation between concentrate class and proportion of trauma bleeding events (dOFV=-250, $p<0.01$). Likewise, covariate analysis on bleed location showed a significant correlation between age and proportion of joint bleeding events (dOFV=-189, $p<0.01$).

The final population RTTE model is presented in the Supplementary Data, including its equations, table of parameter estimates (Table S1), graphical displays of its components (Figures S14 to S19) and evaluations (Figures S20 to S26).

From the RTTE model, the median probability for a bleeding event to happen at a joint location was 53.2, 56.5%, 58.5% and 60.8% for a 6-, 12-, 18- and 30-year-old patient at the start of the study, respectively. However, since the BSV for bleed location was high (CV: 99.6%), the covariate effect may not be predictive or clinically significant.

Evaluations of the final model are provided in the Supplemental Data (Figures S20 to S26). Predicted vs observed overall ABR was well fit with a coefficient of determination of 0.76 (Figure S20). Likewise, cumulative hazard representing the predicted bleed count at time of bleeding event well fit ($R^2$=0.95) the observed bleed count at the same times of event regardless of bleed cause or category (Figure S21). Participants that did not record any bleed (n=137) were predicted with a low bleed count, supported by 111 participants (81%) predicted with two bleeds or less (Figure S22). Stratified VPCs of the Kaplan Meier estimator were also performed to assess if the model was able to capture the observed survival and its variability. Overall, the survival for the occurrence of the first 3 bleeds was well described with observed survival mostly within the 90% CI of the simulated values (Figures S23 and S24) except for the survival for the first bleeding event, which was slightly under-predicted.

Simulations

The comparison between the population RTTE model from literature and the model developed in this study showed very similar hazard and survival over time with wider confidence intervals for the literature model (Figure 4).

The simulation results of Table 5 show a correlation between PK outcomes and bleed risk. Indeed, the higher the Cmin and Cavg, regardless of the dosing regimen or generic patient, the lower the hazard mean value and bleed probabilities. Differences in bleed risks were found between the 4 generic patients, however these differences are correlated to the differences of PK between these generic patients; generic patient (2) – <12 yrs and using an SHL concentrate –

showed the lowest Cmin and Cavg while generic patient (3) – ≥12 yrs and using an EHL concentrate – showed the highest Cmin and Cavg.

## Discussion

This project was initiated to assess whether a better understanding of an individual's pharmacokinetics leads to fewer bleeding events. To this end, real world PK and treatment data available in both the ATHN dataset and the WAPPS-Hemo/CBDR databases were leveraged through PK modeling to simulate FVIII levels along each participant's recorded treatment. Then, bleed data was leveraged in the description of the correlation between factor VIII levels and bleeding risk.

Many studies have investigated potential relationships between PK endpoints (AUC, peak, trough levels, time spent below trough, etc.) and bleeding events [14–18]. These studies showed a correlation between PK endpoints and bleeding events, however this was associated with high inter-individual variability. Previously developed PK-RTTE models also support these results [9,10]. PK-RTTE models provide a parameterized description of the relationship between factor VIII levels and bleeding risk, and therefore can quantify bleeding risk inter-individual variability. The two previously developed PK-RTTE models leveraged clinical trial data of 71 participants using Advate with one year follow up [10] and 183 participants using Kovaltry with one year follow up [9]. The modeling presented in this study aligns with the previous PK-RTTE

models and adds value mainly through its data source and length of observation. The ATHN dataset and WAPPS-Hemo/CBDR databases gather real world data. After curation of the data, PK, treatment and bleed data were available on 427 participants using SHL and EHL concentrates with, on average, 2.3 years follow up. Additionally, knowledge of bleed cause and location was available and allowed a finer tuning of the RTTE model by defining categories of bleeding events. Unfortunately, bleed severity was not available in most of the cases and could not be included in the final model as it was performed in Abrantes et al. [9].

The first component of this study consisted of the Bayesian estimation of individual PK parameters (Table 1). Average and inter-individual variability of PK was consistent with literature and showed significant differences between concentrate classes and age groups [19]. Terminal half-life was, on average, longer in participants using EHL concentrates than in participants using SHL concentrates. Terminal half-life was, on average, longer in participants older than 12 compared to participants younger than 12 years old.

The second component of this study assessed the distributions of simulated FVIII activity at times of bleeding events and supported the correlation observed between PK and bleeding events (Figures 1 and 2). Most of the bleeding events happened at a simulated FVIII < 10 IU/dL with a median [IQR] of 2.62 [0.74 – 9.62] IU/dL. The results were consistent with the median [IQR] of 3.43 [1.33–8.51] IU/dL reported by Valentino *et al*. [18] assessing the distribution of predicted FVIII activity at time of joint bleeding events for 34 participants infused with Advate. The analysis also showed there was no significant difference in predicted FVIII activity between spontaneous (n=57) or traumatic (n=74) bleeding events. However, no significant difference

was found between joint (n=121) and non-joint (n=10) bleeding events, likely due to a low count in non-joint bleeding events (n=131 against 2862 bleeding events in our study). Our study also assessed the difference between bleeding events from participants infused with SHL concentrates against participants infused with EHL concentrates. Simulated FVIII activity at time of bleeding event for participants using EHL concentrates was significantly higher than for participants using SHL concentrates. However, this statistically significant difference was likely due to the lower number of bleeding events for participants using EHL concentrates (n=538 for EHL against for 2324 for SHL) associated with higher average FVIII activity (Figures S6 and S7). The third component of this study assessed the relationship between simulated FVIII activity at times of bleeding events by the development and evaluation of a PK-RTTE model. The final model included time using a Weibull function and FVIII activity using a Hill function. Although, time and FVIII activity modules were not identical to the literature PK-RTTE models [9,10], they show similar patterns especially when ABR is simulated for the typical patient with constant FVIII activity. When constant activity is 1 IU/dL, bleed rates are 4.32, 2.06 and 2.32 bleeds/year for Titman *et al* [10], Abrantes *et al* [9] and the model developed in this study, respectively. When constant activity is 10 IU/dL, bleed rates decrease to 1.05, 1.14 and 1.41 bleeds/year for Titman *et al* [10], Abrantes *et al* [9]. and the model developed in this study, respectively. A time trend was identified in the hazard with a typical patient having a bleed rate 3.6 times higher at the start of the study compared to 2 years after the start. A similar trend was also described by Abrantes *et al.* [9] with a ratio of 3.1 between the hazard values at the start of the study and 2 years after the start. The decrease of bleed rate over time was attributed to treatment effect,

not explained by plasma FVIII activity. It was assumed as either a consequence of a normalization of the clotting system due to prophylactic treatment or a change in adherence to treatment and recording over time.

Covariates were investigated and led to significant effects of concentrate class on bleed cause: traumatic cause was more prevalent in EHL concentrates. However, there was no trend between spontaneous/traumatic cause in SHL concentrates. Although the relationship between the use of EHL concentrates and an increased prevalence in traumatic bleeding events was found to be statistically significant, further work would be required to determine if this relationship has any clinical basis. A significant effect was found for age on bleed location: older patients were more likely to have bleeds at joints. Interestingly, Abrantes *et al.* [9] identified age as a statistically significant covariate of bleed severity, however this was not available in our dataset. No covariate was found significant on bleed risk, which was consistent with literature models [9,10] whose only identified significant covariate was bleed history, also not available in our study.

Standard evaluation methods, including VPC of the Kaplan Meier estimator, demonstrated that the final PK-RTTE model described the bleed data with little bias and reasonable precision. Simulations of virtual doses, intervals, patient covariates are advised to remain within the range defined by the derivation data as the model predictions on an external population have not been evaluated. However, the model here well described the Abrantes simulations (Figure 4) and showed less variability as CV was 111% vs. 79% despite relying on real world data vs. clinical trial data.

Limitations were identified in this study mainly originating from the data source. Since this was a retrospective study that leveraged real world data, important covariates such as bleed history and bleed severity could not be collected and included in the analysis. Such covariates may have reduced the high remaining inter-patient variability of the final model. Moreover, patient adherence may have been lower than in clinical trial settings potentially decreasing reliability of recorded doses and bleeding events. This could have also impacted the precision of the model. However, this model was consistent in terms of estimates and their uncertainty compared to literature models derived from clinical trial data. Consequently, the precision of the model was assumed appropriate and not impacted differently by adherence compared to previous models. We identified a lack of fit in the VPC of the Kaplan Meier estimator for the occurrence of the first bleeding event. This lack of fit was related to the high ratio of non-bleeders in the population (32% of the participants did not record any bleed - Table 2). The lack of fit disappears at a wider scale comparing the VPCs of the occurrence of subsequent bleeds. Model refinement using bleed history as a covariate could also improve the fit of the VPC for the occurrence of the first bleeding event.

This model was developed for the purpose of describing the relationship between bleeding risk and factor VIII activity at the population level. As such, virtual patients and model simulations were kept within the covariate range delimited by the observed data. The model supports the benefits of prophylaxis showing that higher FVIII levels leads to lower bleed risk. Especially, the simulations highlighted that Cmin, Cavg and time spent below specific troughs were well correlated with bleed risk independently from dosing. While Cavg is a good marker, it may not

be a reasonable target for prophylaxis from a clinical standpoint as its estimation requires more than one sample whereas Cmin is measurable and knowable from one sample [18].

Future directions of this work include the external evaluation of the RTTE model in predicting individual ABR using Bayesian forecasting. Internal evaluations already showed promising results regarding the Bayesian forecasting of the bleeding probability performed with Abrantes *et al.* RTTE model [20].

In summary, the model suggests that knowledge of an individual's PK allows for an assessment of important PK outcomes – especially Cmin – that are correlated with bleeding risk. Once PK is individually known, bleeding risk was not further affected by dose, age, BMI nor concentrate class.

## Acknowledgements

This work was supported by an unrestricted educational grant from Sanofi to the American Thrombosis and Hemostasis Network.

## References

1 Shih M, Wang J, Yin J, *et al.* Differences in Major Bleeding Events Between Patients With Severe Hemophilia A and Hemophilia B: A Nationwide, Population-Based Cohort Study. *Clin Appl Thromb Hemost* 2019;**25**:1076029619888023.

2 Manco-Johnson MJ, Abshire TC, Shapiro AD, *et al.* Prophylaxis versus episodic treatment to prevent joint disease in boys with severe hemophilia. *N Engl J Med* 2007;**357**:535–44.

3 Mahlangu J, Oldenburg J, Paz-Priel I, *et al.* Emicizumab Prophylaxis in Patients Who Have Hemophilia A without Inhibitors. *N Engl J Med* 2018;**379**:811–22.

4 Berntorp E, Hermans C, Solms A, *et al.* Optimising prophylaxis in haemophilia A: The ups and downs of treatment. *Blood Reviews* 2021;**50**:100852.

5 Hajducek D, Chelle P, Hermans C, *et al.* Development and evaluation of the population pharmacokinetic models for FVIII and FIX concentrates of the WAPPS-Hemo project. *Haemophilia* 2020;**00**:1–17. doi:10.1111/hae.13977

6 Bergstrand M, Hooker AC, Wallin JE, *et al.* Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J* 2011;**13**:143–51. doi:10.1208/s12248-011-9255-z

7 Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* 2018;**4**:171085.

8 Keizer R. http://vpc.ronkeizer.com/. 2021.

9 Abrantes JA, Solms A, Garmann D, *et al.* Relationship between factor VIII activity, bleeds and individual characteristics in severe hemophilia A patients. *Haematologica* 2019;**104**.

10 Titman A, Wolfsegger M, Jaki T. Recurrent events modelling of haemophilia bleeding events. *J R Stat Soc Series C* 2021;**70**:351–71.

11 Yoneyama K, Schmitt C, Kotani N, *et al.* A Pharmacometric Approach to Substitute for a Conventional Dose-Finding Study in Rare Diseases: Example of Phase III Dose Selection for Emicizumab in Hemophilia A. *Clin Pharmacokinet* 2018;**57**:1123–34.

12 Holford N. A Time to Event Tutorial for Pharmacometricians. *CPT: Pharmacometrics & Systems Pharmacology* 2013;**2**.

13 Savic R, Karlsson M. Importance of Shrinkage in Empirical Bayes Estimates for Diagnostics: Problems and Solutions. *AAPS J* 2009;**11**:558–69. doi:10.1208/s12248-009-9133-0

14 Ahnström J, Berntorp E, Lindvall K, *et al.* A 6-year follow-up of dosing, coagulation factor levels and bleedings in relation to joint status in the prophylactic treatment of haemophilia. *Haemophilia* 2004;**10**:689–97. doi:10.1111/j.1365-2516.2004.01036.x

15 Zhou JY, Barnes RFW, Foster G, *et al.* Joint Bleeding Tendencies in Adult Patients With Hemophilia: It's Not All Pharmacokinetics. *Clinical and Applied Thrombosis/Hemostasis* 2019;**25**:1–10.

16 Broderick CR, Herbert RD, Latimer J, *et al.* Association Between Physical Activity and Risk of Bleeding in Children With Hemophilia. *JAMA* 2012;**308**:1452–1459. doi:10.1001/jama.2012.12727

17 Cheng X, Li P, Chen Z, *et al.* Break-through bleeding in relation to pharmacokinetics of Factor VIII in paediatric patients with severe haemophilia A. *Haemophilia* 2018;**24**:120–5. doi:10.1111/hae.13373

18 Valentino L, Pipe S, Collins P, *et al.* Association of peak factor VIII levels and area under the curve with bleeding in patients with haemophilia A on every third day pharmacokinetic-guided prophylaxis. *Haemophilia* 2016;**22**:514–20. doi:10.1111/hae.12905

19 Versloot O, Iserman E, Chelle P, *et al.* Terminal half-life of FVIII and FIX according to age, blood group and concentrate type: Data from the WAPPS database. *J Thromb Haemost* 2021;**19**:1896–1906. doi:10.1111/jth.15395

20 Abrantes JA, Solms A, Garmann D, *et al.* Bayesian Forecasting Utilizing Bleeding Information to Support Dose Individualization of Factor VIII. *CPT: Pharmacometrics & Systems Pharmacology* 2019;**8**:894–903. doi:10.1002/psp4.12464

# Tables

Table 1: Summary of study participant characteristics and their estimated FVIII PK profiles. Data are summarized as median [min-max]

| Parameter | Concentrate | Adolescents and Adults (≥ 12yrs) | Children (< 12 yrs) | All |
|---|---|---|---|---|
| Patient characteristics | | | | |
| Age (yrs) | EHL | (n=131) 28.75 [12-66.1] | (n=37) 7.92 [1-11.83] | (n=168) 21.7 [1-66.1] |
| | SHL | (n=180) 27.71 [12-68] | (n=88) 5.96 [0.75-11.9] | (n=268) 20 [0.75-68] |
| | All | (n=311) 28.17 [12-68] | (n=125) 6.00 [0.75-11.9] | (n=436) 20.5 [0.75-68] |
| Weight (kg) | EHL | (n=131) 74.5 [29-150] | (n=37) 26.9 [8.4-59.4] | (n=168) 69.15 [8.4-150] |
| | SHL | (n=180) 78.2 [33.2-143] | (n=88) 21.05 [7.4-57] | (n=268) 67.3 [7.4-143] |
| | All | (n=311) 76.7 [29-150] | (n=125) 21.6 [7.4-59.4] | (n=436) 68 [7.4-150] |
| % Severe (Endogenous FVIII <1 IU/dL) | EHL | (n=112/131) 85.5 % | (n=29/37) 78.4 % | (n=141/168) 83.9 % |
| | SHL | (n=143/180) 79.4 % | (n=82/88) 93.2 % | (n=225/268) 84.0 % |
| | All | (n=255/311) 82.0 % | (n=111/125) 88.8 % | (n=366/436) 83.9 % |
| % Positive History of Inhibitors | EHL | (n=4/131) 3.05 % | (n=12/37) 32.4 % | (n=16/168) 9.52 % |
| | SHL | (n=16/180) 8.89 % | (n=22/88) 25.0 % | (n=38/268) 14.2 % |
| | All | (n=20/311) 6.43 % | (n=34/125) 27.2 % | (n=54/436) 12.4 % |
| FVIII PK | | | | |
| Dose (IU/kg) | EHL | (n=131) 36.15 [11.32-94.51] | (n=37) 50 [28.34-148.32] | (n=168) 38.45 [11.32-148.3] |
| | SHL | (n=180) 29.48 [8.92-89.53] | (n=88) 42.29 [17.06-200] | (n=268) 30.97 [8.92-200] |
| | All | (n=311) 31.02 [8.92-94.51] | (n=125) 44.44 [17.06-200] | (n=436) 34.01 [8.92-200] |
| Number of Samples per infusion | EHL | (n=131) 4 [1-7] | (n=37) 3 [1-11] | (n=168) 4 [1-11] |
| | SHL | (n=180) 3 [1-21] | (n=88) 3 [1-14] | (n=268) 3 [1-21] |

| | | | | |
|---|---|---|---|---|
| | All | (n=311) 4 [1-21] | (n=125) 3 [1-14] | (n=436) 3 [1-21] |
| Terminal Half-life (h) | EHL | (n=131) 16 [5.2-26.8] | (n=37) 13.1 [8.6-19.5] | (n=168) 14.65 [5.2-26.8] |
| | SHL | (n=180) 11.55 [5-36.1] | (n=88) 8.45 [2.2-17.5] | (n=268) 10.7 [2.2-36.1] |
| | All | (n=311) 13.3 [5-36.1] | (n=125) 10.1 [2.2-19.5] | (n=436) 12.2 [2.2-36.1] |
| Clearance (mL/h/kg) | EHL | (n=131) 2.11 [0.87-5.71] | (n=37) 2.91 [2.25-5.27] | (n=168) 2.29 [0.87-5.71] |
| | SHL | (n=180) 3.12 [0.72-21.5] | (n=88) 5.11 [1.88-32.56] | (n=268) 3.54 [0.72-32.56] |
| | All | (n=311) 2.61 [0.72-21.5] | (n=125) 4.55 [1.88-32.56] | (n=436) 2.85 [0.72-32.56] |
| Volume at steady state (mL/kg) | EHL | (n=131) 46.04 [21.77-78.7] | (n=37) 57.98 [37-81.77] | (n=168) 48.02 [21.77-81.77] |
| | SHL | (n=180) 48.95 [26.05-160.17] | (n=88) 62.01 [36.87-131.48] | (n=268) 52.35 [26.05-160.17] |
| | All | (n=311) 47.82 [21.77-160.17] | (n=125) 60.84 [36.87-131.48] | (n=436) 50.09 [21.77-160.17] |
| Simulated Time to 1 IU/dL from a 50 IU/kg dose (h) | EHL | (n=112) 121.94 [41.36-230.73] | (n=29) 93.75 [56.73-137.48] | (n=141) 113.8 [41.4-230.7] |
| | SHL | (n=143) 83.32 [31.25-282.18] | (n=82) 62.48 [13.95-123.18] | (n=225) 77.1 [13.95-282.18] |
| | All | (n=255) 97 [31.25-282.18] | (n=111) 69.48 [13.95-137.48] | (n=366) 90.5 [13.95-282.18] |
| Simulated Time to 3 IU/dL from a 50 IU/kg dose (h) | EHL | (n=126) 85.42 [29.21-154.71] | (n=36) 65.55 [40.52-111.18] | (n=162) 81 [29.21-154.71] |
| | SHL | (n=161) 58.29 [18.68-198.73] | (n=87) 42.69 [8.83-85.39] | (n=248) 52.62 [8.83-198.73] |
| | All | (n=287) 70.21 [18.68-198.73] | (n=123) 49.23 [8.83-111.18] | (n=410) 63.89 [8.83-198.73] |
| Simulated Time to 5 IU/dL from a 50 IU/kg dose (h) | EHL | (n=130) 71.98 [24.77-123.66] | (n=37) 55.97 [33.7-83.31] | (n=167) 67.58 [24.77-123.66] |
| | SHL | (n=165) 47.95 [14.2-168.25] | (n=88) 35.28 [6.98-71.75] | (n=253) 43.99 [6.98-168.25] |
| | All | (n=295) 58.59 [14.2-168.25] | (n=125) 40.78 [6.98-83.31] | (n=420) 53.75 [6.98-168.25] |

Table 2: Summary of study dosing and bleed characteristics. Data are summarized as median [min-max]

| Parameter | Concentrate | Adolescents and Adults (≥ 12yrs) | Children (< 12 yrs) | All |
|---|---|---|---|---|
| Observation and Dosing characteristics | | | | |
| Observation Period (yrs) | EHL | (n=130) 0.76 [0.04-11.22] | (n=35) 1.36 [0.06-7.36] | (n=165) 1.13 [0.04-11.22] |
| | SHL | (n=182) 2.7 [0.07-11.22] | (n=80) 2.83 [0.13-10.82] | (n=262) 2.71 [0.07-11.22] |
| | All | (n=312) 2.34 [0.04-11.22] | (n=115) 2 [0.06-10.82] | (n=427) 2.28 [0.04-11.22] |
| Number of Recorded Doses | EHL | (n=130) 68 [3-1537] | (n=35) 134 [4-649] | (n=165) 72 [3-1537] |
| | SHL | (n=182) 288.5 [4-1649] | (n=80) 250.5 [3-1313] | (n=262) 283 [3-1649] |
| | All | (n=312) 168.5 [3-1649] | (n=115) 183 [3-1313] | (n=427) 175 [3-1649] |
| Median Dose (IU/kg) | EHL | (n=130) 34.8 [10-88.69] | (n=35) 50.34 [16.84-123.93] | (n=165) 37.31 [10-123.93] |
| | SHL | (n=182) 29.85 [7.04-94.45] | (n=80) 43.67 [17.06-321.43] | (n=262) 32.04 [7.04-321.43] |
| | All | (n=312) 31.86 [7.04-94.45] | (n=115) 46.73 [16.84-321.43] | (n=427) 34.31 [7.04-321.43] |
| % On demand (Median Dose Interval > 14 days) | EHL | (n=2/130) 1.54 % | (n=1/35) 2.86 % | (n=3/165) 1.82 % |
| | SHL | (n=2/182) 1.10 % | (n=4/80) 5.00 % | (n=6/262) 2.29 % |
| | All | (n=4/312) 1.28 % | (n=5/115) 4.35 % | (n=9/427) 2.11 % |
| Median Dose Interval (days) | EHL | (n=130) 3.02 [1-96.06] | (n=35) 3 [1-123] | (n=165) 3 [1-123] |
| | SHL | (n=182) 2.14 [0.5-395.2] | (n=80) 2.02 [0.34-218] | (n=262) 2.06 [0.34-395.2] |
| | All | (n=312) 2.98 [0.5-395.2] | (n=115) 2.27 [0.34-218] | (n=427) 2.96 [0.34-395.2] |
| Bleed characteristics | | | | |
| Total Number of Bleeds | EHL | (n=130) 1 [0-62] | (n=35) 1 [0-8] | (n=165) 1 [0-62] |
| | SHL | (n=182) 4 [0-199] | (n=80) 3 [0-93] | (n=262) 4 [0-199] |
| | All | (n=312) 2 [0-199] | (n=115) 2 [0-93] | (n=427) 2 [0-199] |
| Annual Bleed Rate (bleeds/yr) | EHL | (n=130) 0.3 [0-37.05] | (n=35) 0.76 [0-31.7] | (n=165) 0.41 [0-37.05] |
| | SHL | (n=182) 1.41 [0-33.04] | (n=80) 1.67 [0-15.41] | (n=262) 1.47 [0-33.04] |
| | All | (n=312) 1.04 [0-37.05] | (n=115) 1.34 [0-31.7] | (n=427) 1.1 [0-37.05] |

| | | | | |
|---|---|---|---|---|
| Annual Joint Bleed Rate (bleeds/yr) | EHL | (n=130) 0 [0-37.05] | (n=35) 0 [0-5.13] | (n=165) 0 [0-37.05] |
| | SHL | (n=182) 0.8 [0-25.2] | (n=80) 0.37 [0-6.91] | (n=262) 0.67 [0-25.2] |
| | All | (n=312) 0.57 [0-37.05] | (n=115) 0.34 [0-6.91] | (n=427) 0.39 [0-37.05] |
| Annual Non-joint Bleed Rate (bleeds/yr) | EHL | (n=130) 0 [0-7.43] | (n=35) 0 [0-31.7] | (n=165) 0 [0-31.7] |
| | SHL | (n=182) 0.35 [0-12.02] | (n=80) 0.48 [0-15.41] | (n=262) 0.36 [0-15.41] |
| | All | (n=312) 0 [0-12.02] | (n=115) 0.39 [0-31.7] | (n=427) 0.19 [0-31.7] |
| Annual Spontaneous Bleed Rate (bleeds/yr) | EHL | (n=130) 0 [0-37.05] | (n=35) 0 [0-3.22] | (n=165) 0 [0-37.05] |
| | SHL | (n=182) 0.38 [0-20.6] | (n=80) 0.35 [0-8.06] | (n=262) 0.37 [0-20.6] |
| | All | (n=312) 0.32 [0-37.05] | (n=115) 0 [0-8.06] | (n=427) 0 [0-37.05] |
| Annual Trauma Bleed Rate (bleeds/yr) | EHL | (n=130) 0 [0-14.86] | (n=35) 0.46 [0-31.7] | (n=165) 0 [0-31.7] |
| | SHL | (n=182) 0.67 [0-22.4] | (n=80) 0.79 [0-15.41] | (n=262) 0.7 [0-22.4] |
| | All | (n=312) 0.35 [0-22.4] | (n=115) 0.71 [0-31.7] | (n=427) 0.43 [0-31.7] |
| % No bleed | EHL | (n=63/130) 48.5 % | (n=14/35) 40.0 % | (n=77/165) 46.7 % |
| | SHL | (n=46/182) 25.3 % | (n=14/80) 17.5 % | (n=60/262) 22.9 % |
| | All | (n=109/312) 34.9 % | (n=28/115) 24.3 % | (n=137/427) 32.1 % |
| % No spontaneous bleed | EHL | (n=77/130) 59.2 % | (n=30/35) 85.7 % | (n=107/165) 64.8 % |
| | SHL | (n=71/182) 39.0 % | (n=36/80) 45.0 % | (n=107/262) 40.8 % |
| | All | (n=148/312) 47.4 % | (n=66/115) 57.4 % | (n=214/427) 50.1 % |
| % No joint bleed | EHL | (n=70/130) 53.8 % | (n=22/35) 62.9 % | (n=92/165) 55.8 % |
| | SHL | (n=63/182) 34.6 % | (n=25/80) 31.3 % | (n=88/262) 33.6 % |
| | All | (n=133/312) 42.6 % | (n=47/115) 40.9 % | (n=180/427) 42.2 % |

Table 3: Matrix of bleed categories

| Bleed Cause / Bleed Location | Trauma | Spontaneous bleed | All causes |
|---|---|---|---|
| Joint | 854 (29.8 %) | 1040 (36.3 %) | 1894 (66.2 %) |
| No Joint | 580 (20.3 %) | 388 (13.6 %) | 968 (33.8 %) |
| All locations | 1434 (50.1 %) | 1428 (49.9 %) | 2862 (100 %) |

Table 4: Results of simulations comparing the effect of prophylactic treatment dose and interval on PK and bleed risk (mean hazard, bleed probability within 6 months and within 1 year) between 4 generic patients - (1) Adolescents/Adults infused with an SHL concentrate, (2) Children infused with an SHL concentrate, (3) Adolescents/Adults infused with an EHL concentrate, and (4) Children infused with an EHL concentrate.

| **Concentrate class** | **Age group** | **Simulated Dose (IU/kg)** | **Simulated Interval (days)** | **Cmin (IU/dL)** | **Cavg (IU/dL)** | **TAT 3 (%)** | **Hazard Mean Value (bleeds/yr)** | | | **Bleed Probability within 6 months (%)** | | | **Bleed Probability within 1 year (%)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Typical | 5th Percentile | 95th Percentile | Typical | 5th Percentile | 95th Percentile | Typical | 5th Percentile | 95th Percentile |
| SHL | Adolescents / Adults (≥ 12yrs) | 35 | 3 | 1.7 | 18.2 | 80 | 1.21 | 0.33 | 4.4 | 55.3 | 15.6 | 97.8 | 76 | 25.9 | 99.9 |
| SHL | Children (< 12 yrs) | 35 | 3 | 0.7 | 11.1 | 53.4 | 1.47 | 0.4 | 5.37 | 62.7 | 18.7 | 99.1 | 82.6 | 30.7 | 100 |
| EHL | Adolescents / Adults (≥ 12yrs) | 35 | 3 | 3.8 | 25.6 | 100 | 0.98 | 0.27 | 3.58 | 47.8 | 12.8 | 95.5 | 68.4 | 21.5 | 99.6 |
| EHL | Children (< 12 yrs) | 35 | 3 | 1.6 | 16.8 | 76.7 | 1.24 | 0.34 | 4.51 | 56.2 | 15.9 | 98 | 76.8 | 26.4 | 99.9 |
| SHL | Adolescents / Adults (≥ 12yrs) | 50 | 3 | 2.2 | 25.8 | 90 | 1.07 | 0.29 | 3.92 | 51.1 | 13.9 | 96.7 | 71.8 | 23.4 | 99.8 |
| SHL | Children | 50 | 3 | 0.8 | 15.7 | 60.1 | 1.37 | 0.38 | 5 | 60.1 | 17.5 | 98.7 | 80.4 | 28.9 | 100 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (< 12 yrs) | | | | | | | | | | | | | |
| EHL | Adolescents / Adults (≥ 12yrs) | 50 | 3 | 5.2 | 36.3 | 100 | 0.83 | 0.23 | 3.04 | 42.5 | 11 | 92.8 | 62.4 | 18.6 | 99.1 |
| EHL | Children (< 12 yrs) | 50 | 3 | 2 | 23.8 | 86.7 | 1.1 | 0.3 | 4.03 | 52.1 | 14.3 | 97 | 72.9 | 24 | 99.8 |
| SHL | Adolescents / Adults (≥ 12yrs) | 20 | 3 | 1.2 | 10.6 | 66.7 | 1.4 | 0.38 | 5.11 | 60.9 | 17.9 | 98.9 | 81 | 29.5 | 100 |
| SHL | Children (< 12 yrs) | 20 | 3 | 0.6 | 6.6 | 43.4 | 1.61 | 0.44 | 5.88 | 66.2 | 20.3 | 99.4 | 85.3 | 33.2 | 100 |
| EHL | Adolescents / Adults (≥ 12yrs) | 20 | 3 | 2.4 | 14.8 | 90 | 1.21 | 0.33 | 4.41 | 55.4 | 15.6 | 97.9 | 76 | 25.9 | 99.9 |
| EHL | Children (< 12 yrs) | 20 | 3 | 1.1 | 9.8 | 63.4 | 1.43 | 0.39 | 5.21 | 61.6 | 18.2 | 99 | 81.6 | 29.9 | 100 |
| SHL | Adolescents / Adults (≥ 12yrs) | 35 | 2 | 5.5 | 27 | 100 | 0.92 | 0.25 | 3.34 | 45.6 | 12 | 94.5 | 65.9 | 20.2 | 99.4 |
| SHL | Children (< 12 yrs) | 35 | 2 | 1.8 | 16.5 | 80 | 1.22 | 0.34 | 4.47 | 55.9 | 15.8 | 98 | 76.5 | 26.2 | 99.9 |
| EHL | Adolescents / Adults (≥ 12yrs) | 35 | 2 | 11.1 | 38.1 | 100 | 0.7 | 0.19 | 2.54 | 36.9 | 9.2 | 88.8 | 55.7 | 15.7 | 97.9 |
| EHL | Children (< 12 yrs) | 35 | 2 | 5 | 24.9 | 100 | 0.95 | 0.26 | 3.47 | 46.8 | 12.4 | 95.1 | 67.3 | 20.9 | 99.5 |
| SHL | Adolescents / | 35 | 4 | 0.8 | 13.8 | 60.1 | 1.39 | 0.38 | 5.08 | 60.6 | 17.8 | 98.8 | 80.8 | 29.3 | 100 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Adults (≥ 12yrs) | | | | | | | | | | | | | | |
| SHL | Children (< 12 yrs) | 35 | 4 | 0.5 | 8.5 | 40.1 | 1.6 | 0.44 | 5.86 | 66 | 20.3 | 99.4 | 85.2 | 33 | 100 |
| EHL | Adolescents / Adults (≥ 12yrs) | 35 | 4 | 1.6 | 19.3 | 80.1 | 1.19 | 0.33 | 4.35 | 54.9 | 15.4 | 97.7 | 75.6 | 25.6 | 99.9 |
| EHL | Children (< 12 yrs) | 35 | 4 | 0.8 | 12.7 | 57.6 | 1.42 | 0.39 | 5.17 | 61.3 | 18.1 | 98.9 | 81.4 | 29.7 | 100 |

# Figure Legends

Figure 1: Distribution of FVIII activity at times of bleeding events. Vertical red and blue lines highlight values of 1 and 10 IU/dL.

Figure 2: Distribution of FVIII activity at times of bleeding events split by factor concentrate class (SHL, EHL) and bleed cause (trauma, spontaneous). Vertical red and blue lines highlight values of 1 and 10 IU/dL.

Figure 3: Kaplan-Meier curves representing the probabilities of having less than 1 to 9 bleeds during the first 3 years of observation time.

Figure 4: Comparison of hazard and survival between two RTTE models using the same simulated FVIII time-profile activity.
Top left: Simulated FVIII profile using the PK parameters of the generic adult patient infused with an SHL concentrate. Simulated treatment was 35 IU/kg every 3 days during 2 years (Cmin was 1.7 IU/dL).
Top right: Median, $5^{th}$ and $95^{th}$ percentiles of population hazard for developed model (blue) and Abrantes *et al.* model [9] (red).
Bottom left: Median, $5^{th}$ and $95^{th}$ percentiles of population survival for developed model (blue) and Abrantes *et al.* model [9] (red).
Bottom right: Median, $5^{th}$ and $95^{th}$ percentiles of population bleed probability for developed model (blue) and Abrantes *et al.* model [9] (red).

## Figures

Figure 1: Distribution of FVIII activity at times of bleeding events. Vertical red and blue lines highlight values of 1 and 10 IU/dL.

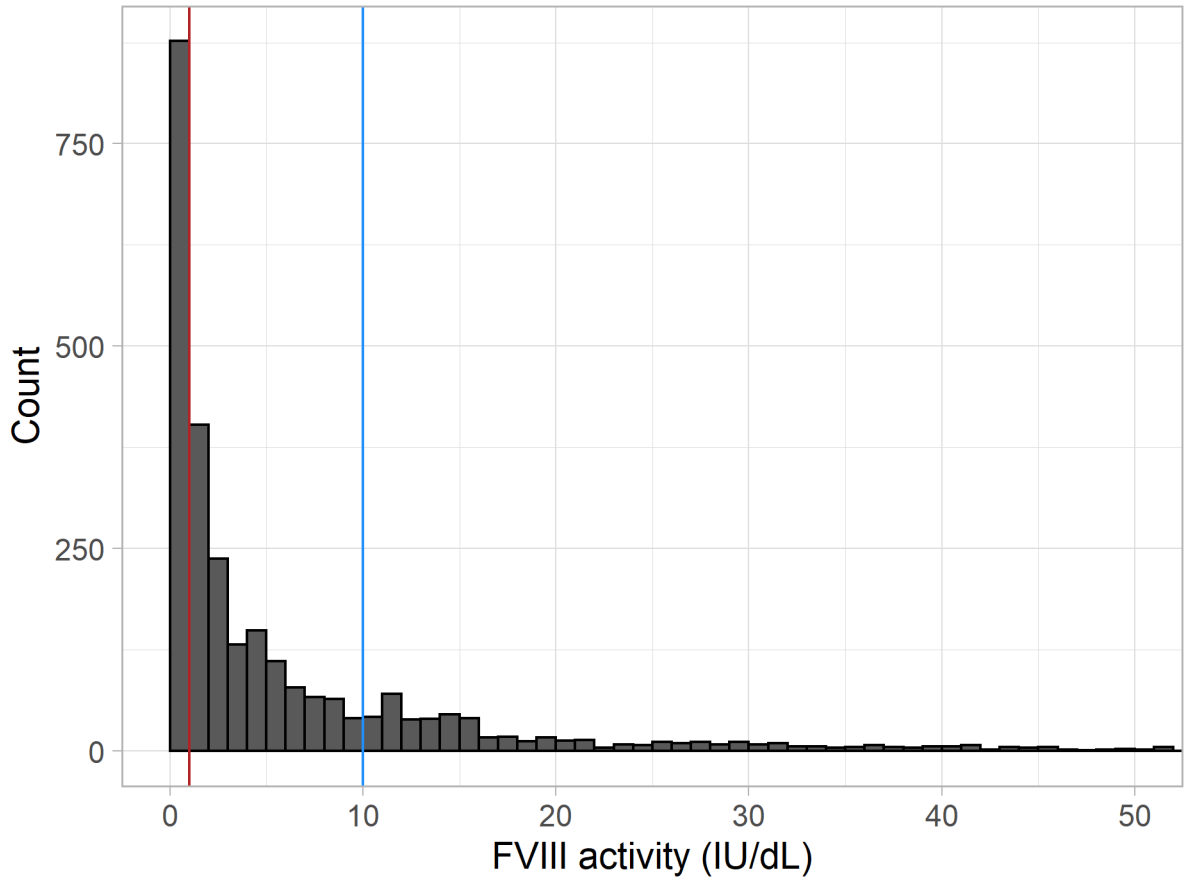


Figure 2: Distribution of FVIII activity at times of bleeding events split by factor concentrate and bleed cause.

Vertical red and blue lines highlight values of 1 and 10 IU/dL.

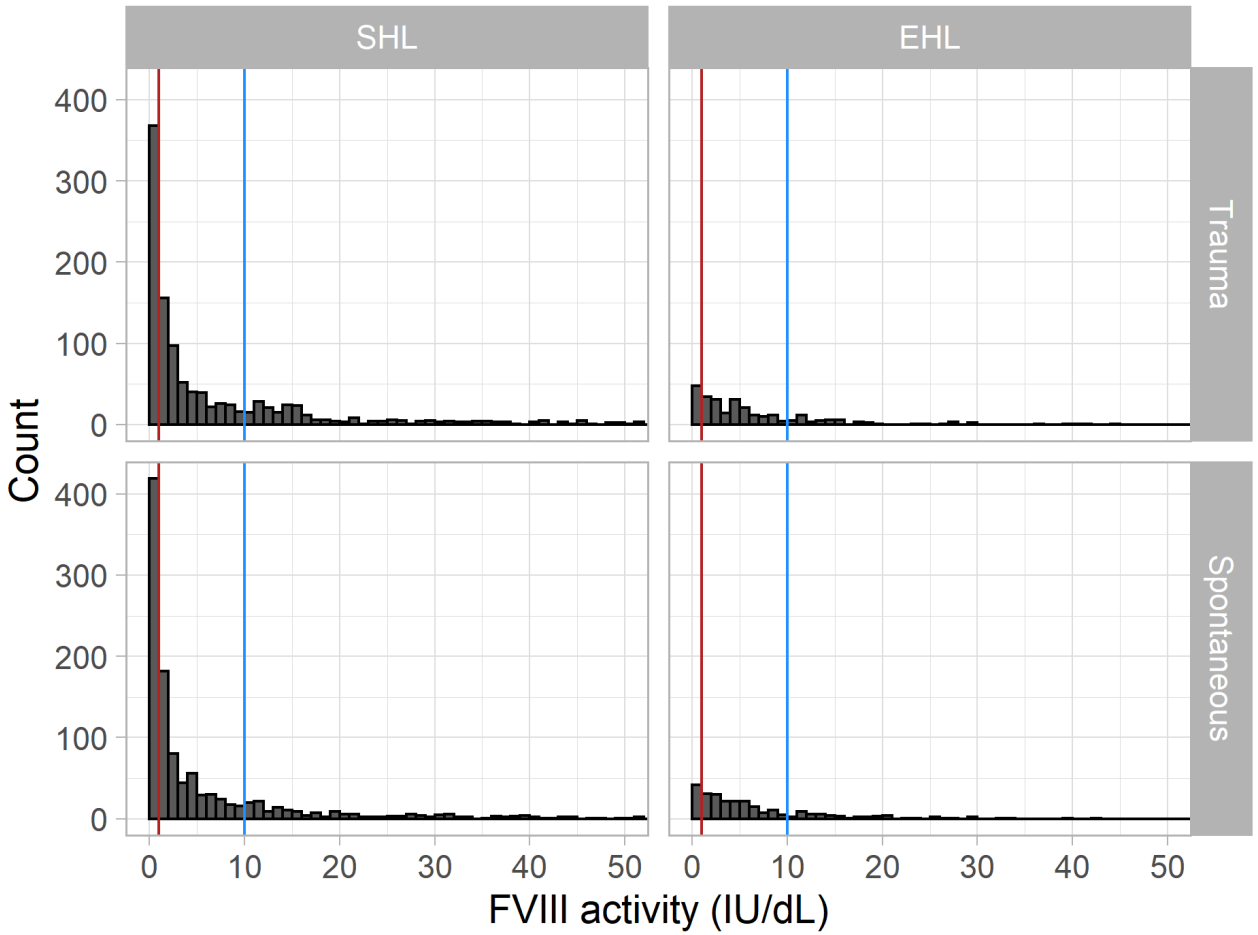

SHL
EHL
Trauma
Spontaneous
Count
400
300
200
100
0
0
10
20
30
40
50
FVIII activity (IU/dL)

Figure 3: Kaplan-Meier curves representing the probabilities of having less than 1 to 9 bleeds during the first 3 years of observation time.

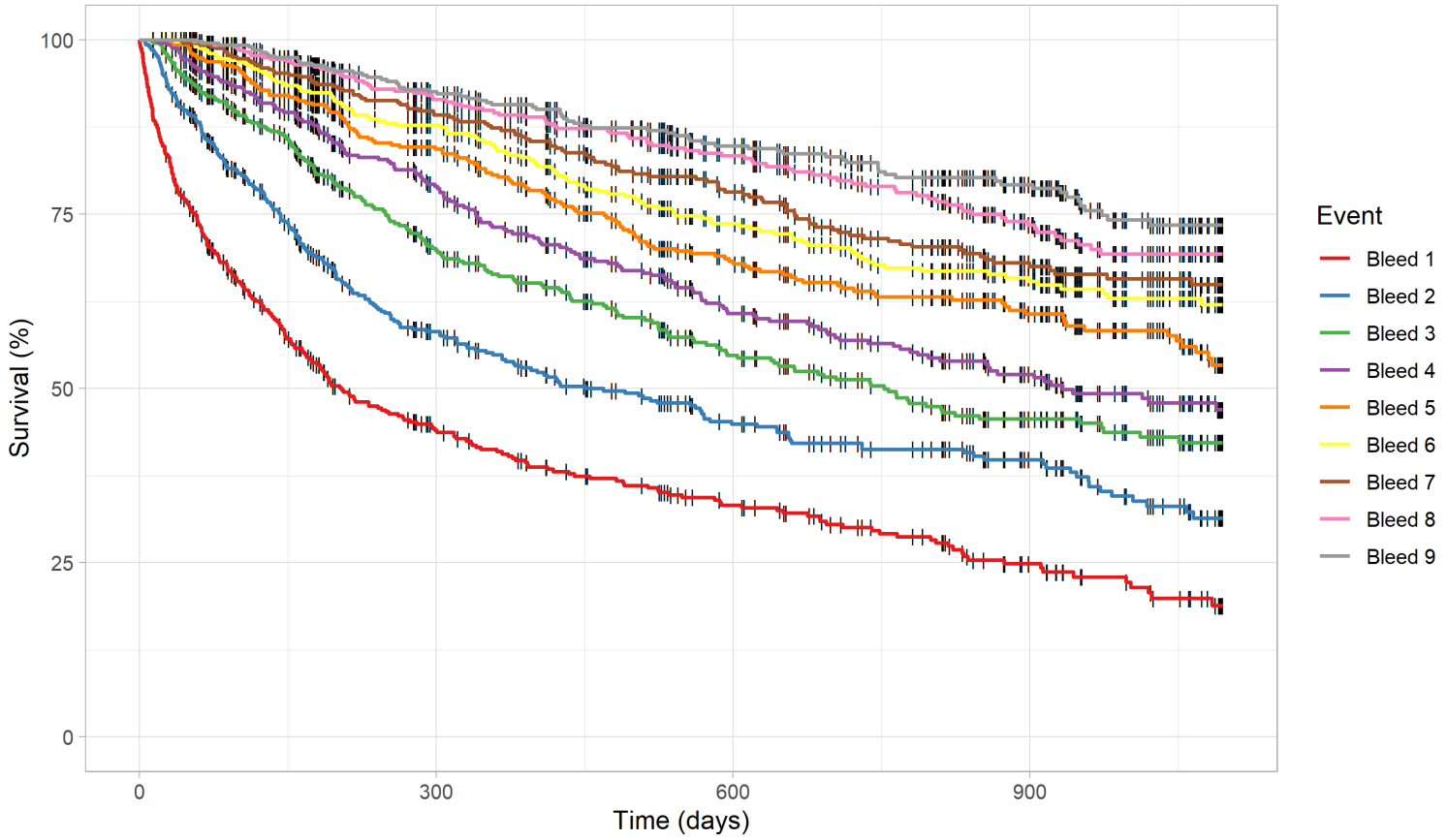


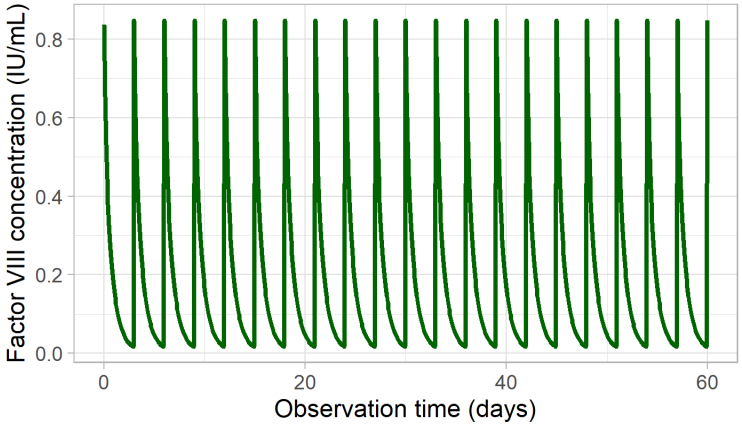
Figure 4: Comparison of hazard and survival between two RTTE models using the same simulated FVIII time-profile activity.

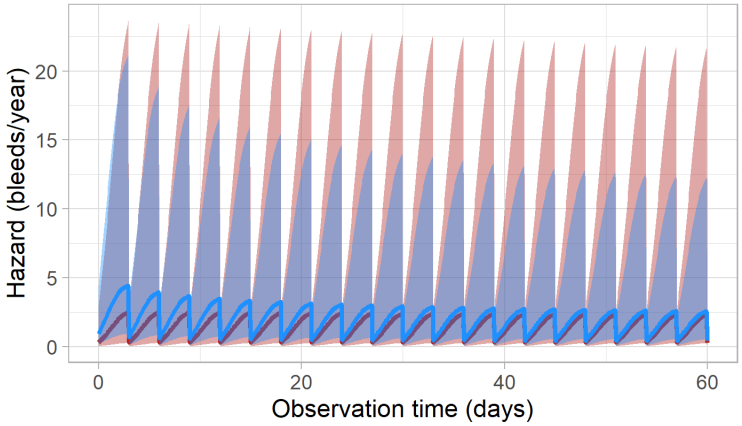
Top left: Simulated FVIII profile using the PK parameters of the generic adult patient infused with an SHL concentrate. Simulated treatment was 35 IU/kg every 3 days during 2 years (Cmin was 1.7 IU/dL).

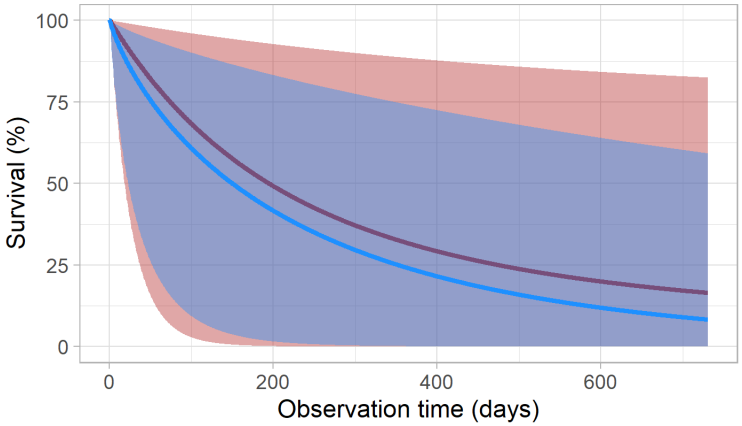
Top right: Median, $5^{th}$ and $95^{th}$ percentiles of population hazard for developed model (blue) and Abrantes *et al.* model [9] (red).

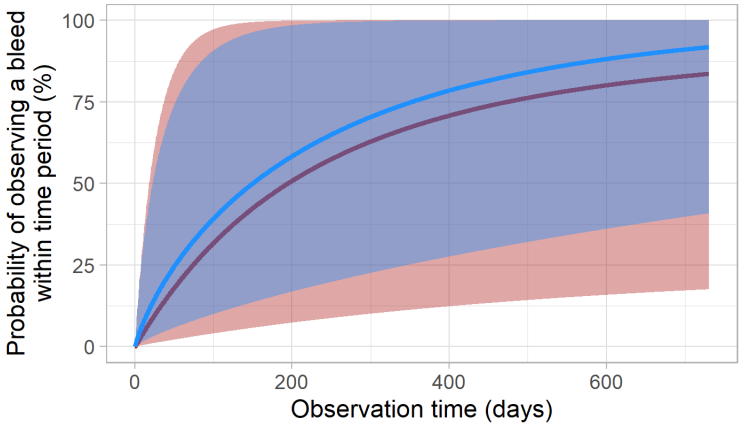
Bottom left: Median, $5^{th}$ and $95^{th}$ percentiles of population survival for developed model (blue) and Abrantes *et al.* model [9] (red).

Bottom right: Median, $5^{th}$ and $95^{th}$ percentiles of population bleed probability for developed model (blue) and Abrantes *et al.* model [9] (red).


Factor VIII concentration (IU/mL)
0.0
0.2
0.4
0.6
0.8
0
20
40
60
Observation time (days)


Hazard (bleeds/year)
0
5
10
15
20
0
20
40
60
Observation time (days)


Survival (%)
0
25
50
75
100
0
200
400
600
Observation time (days)


Probability of observing a bleed within time period (%)
0
25
50
75
100
0
200
400
600
Observation time (days)

# Supplemental data

# Methods

## Repeated time to event population modeling

### Model development and evaluations

Model development corresponded to the selection of the best structural model followed by a covariate analysis. Observation time and simulated FVIII activity were shown to affect hazard in previous RTTE models. The structural model selection corresponded to the inclusion of the models best describing the relationships (i) between observation time and hazard and (ii) between simulated FVIII activity and hazard.

#### *Structural model selection*

The hazard function, $h$, describing the expected bleeding rate at time *t* as assumed with the structure:

$$h(t) = f(t, FVIII(t)) * e^{\eta}$$

where

- *FVIII(t)* is factor VIII activity at time *t,* derived from PK parameters estimated during the previous step and observed doses recorded by each subject
- *f* is the relationship between the hazard function (*h*), time (*t*) and FVIII activity (*FVIII*),
- $\eta$ is the inter-individual deviation from the average hazard.

The survival function $S(t_{int})$ represents the probability of not having a bleed in a time interval [$t_0$ - $t_{int}$]:

$$S(t_{int}) = exp\left(-\int_{t_0}^{t_{int}} h(t)dt\right)$$

where *h(t)* is the hazard function that defines the expected bleeding rate.

Between subject variability (BSV)

BSV of the hazard function – i.e. variability of $\eta$ - was assumed log normal based on literature models. Addition of terms corresponding to the response of hazard to time or/and FVIII activity was investigated.

Relationship between hazard and time

Since observation time may affect hazard, – i.e. bleed events could be more likely to occur at the beginning or at the end of the study - the relationship between time and hazard was tested through the following parametric models:

- Exponential model: $h$ is independent from $t$ (survival profile follows an exponential decay)
- Gompertz model: $h(t) \propto e^{\beta t}$
- Weibull model: $h(t) \propto e^{\beta \ln (t+1)}$

Relationship between hazard and FVIII activity

The relationship between hazard and FVIII activity defines how the FVIII activity at a given moment affects the expected bleeding rate. The relationship between FVIII activity and hazard was tested through the following parametric models:

- No effect: $h$ is independent from FVIII activity
- Hill model: $h(t) \propto 1 - \frac{FVIII}{FVIII+EC50}$

- Power model: $h(t) \propto FVIII^{\beta}$

Residual unexplained variability (RUV)

The estimation of the parameters of the hazard function is performed by maximum likelihood methods which use the likelihood of observing an event when bleed events occur and the likelihood of not observing an event due to censoring at the end of the observation period (i.e., right censoring). The likelihoods of these two possibilities are directly obtained from the survival and hazard functions. The likelihood of not observing an event after a known time due to right censoring is the survival at that time record; and the likelihood of observing an exact event when bleed events occur is equal to the hazard multiplied by survival at a given time. Therefore, no residual unexplained variability was used nor calculated in the RTTE model.

Integration of bleed categories

Bleed categories were included as categories of observed events and not covariates in order to investigate the effect of additional covariates on them. As a consequence, 5 values are possible for the observed dependent variable: no bleed, trauma bleed at joint location, spontaneous bleed at joint location, trauma bleed not at joint location and spontaneous bleed not at joint location.

The probabilities for a bleed depending on their cause (trauma or spontaneous) or location (joint or other) were assumed to follow a logit distribution from a proportional odds model as described in Abrantes *et al*. Such probabilities were defined as follows:

- $P_{trauma} = \frac{e^{tr_0+\eta}}{1+e^{tr_0+\eta}}$
- $P_{spontaneous} = 1 - P_{trauma}$

- $P_{joint} = \frac{e^{j_0+\eta}}{1+e^{j_0+\eta}}$
- $P_{no\ joint} = 1 - P_{joint}$

Where $tr_0$and $j_0$ represent the average effects of the cause (trauma) or location (joint) of the bleed in the log odds (log p/1-p) concerning the respective probabilities; and $\eta$ is the inter-individual deviation from the average.

These probabilities are simultaneously estimated with the hazard and survival functions and used to ponder the probability density function of observing a bleed event.

Model evaluations

Evaluations of the population model were based on its scientific plausibility, changes in objective function values and information criteria (OFV: -2log-likelihood, AIC and BIC), goodness of fit plots and precision of parameter estimates; in accordance with the tutorial [12] and previous published RTTE model of hemophilia bleeds [9]. Between subject variability was evaluated through the assessment of $\eta$ distributions and their shrinkage values. Since the model objective was descriptive rather than predictive, values of $\eta$-shrinkage of up to 50% were considered satisfactory and kept in mind while interpreting $\eta$ vs. covariates plots.

Model evaluations also included dedicated methods for RTTE modeling such as Visual Predictive Checks (VPCs) of the Kaplan-Meier estimator, stratified by occurrence of first, second and third bleed, and stratified by bleed cause/location for the first bleed event. The VPC of the Kaplan-Meier estimator requires the evaluation of the cumulative bleeding hazard function to simulate randomly bleeding time events. These bleeding time events were simulated by using the inverse transform method of the uniform distribution: the

probabilities of occurrence of a bleed ($P$) at times of bleeding events, $t_{bleed}$, are simulated by a random variable ($u$ uniformly distributed between 0 and 1). These probabilities are directly related to the cumulative hazard function ($H$) obtained by integration of hazard over time. Consequently, the times of bleeding events are obtained by solving the equation: $P(t_{bleed}) = 1 - e^{-H(t_{bleed})} = u \leftrightarrow t_{bleed} = H^{-1}(-\log(1-u))$ with $u \in [0,1]$.

Dropout was simulated based on the empirical distribution of the observed dropout times in the data (R *ecdf* function) while accounting for the decrease in population size across time. Some simulated bleeding times were therefore right censored at dropout. When the simulated dropout time exceeded the observed, median dose and dose interval values were assigned to the patient, otherwise the available dosing information was used. From the simulated treatments and FVIII values, hazard and cumulative hazard values were derived as a function of time.

# Results

## PK analysis

Simulated vs observed factor activities were found in satisfactory agreement ($R^2$=94%), ensuing the reliable use of this component in the model (Figure S1).

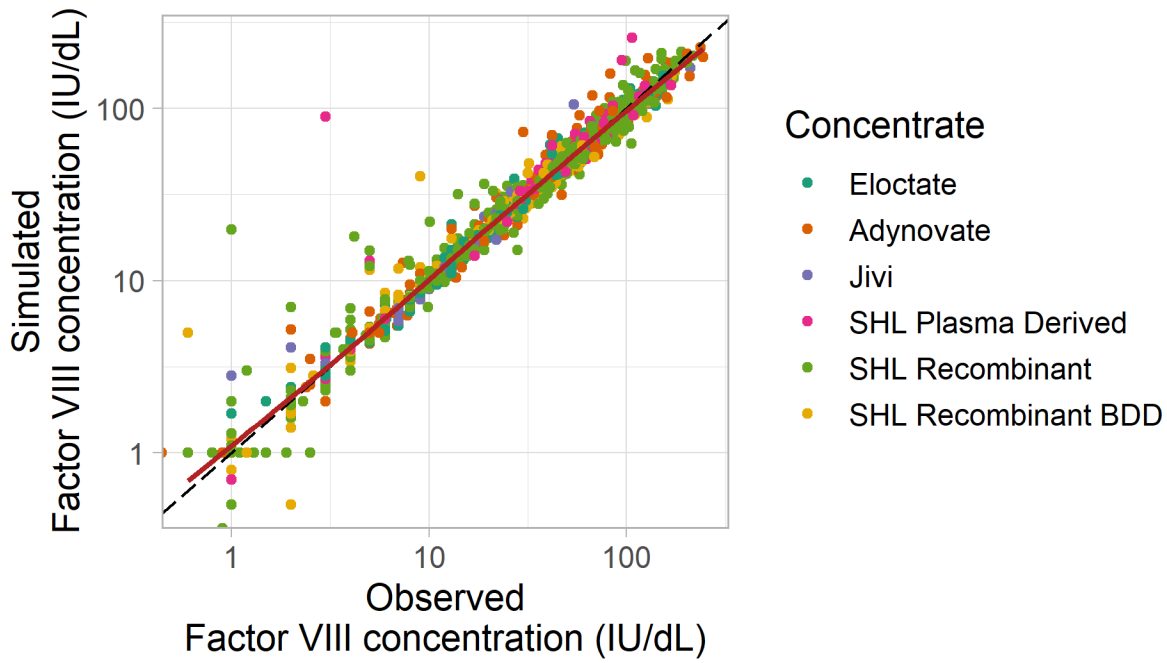
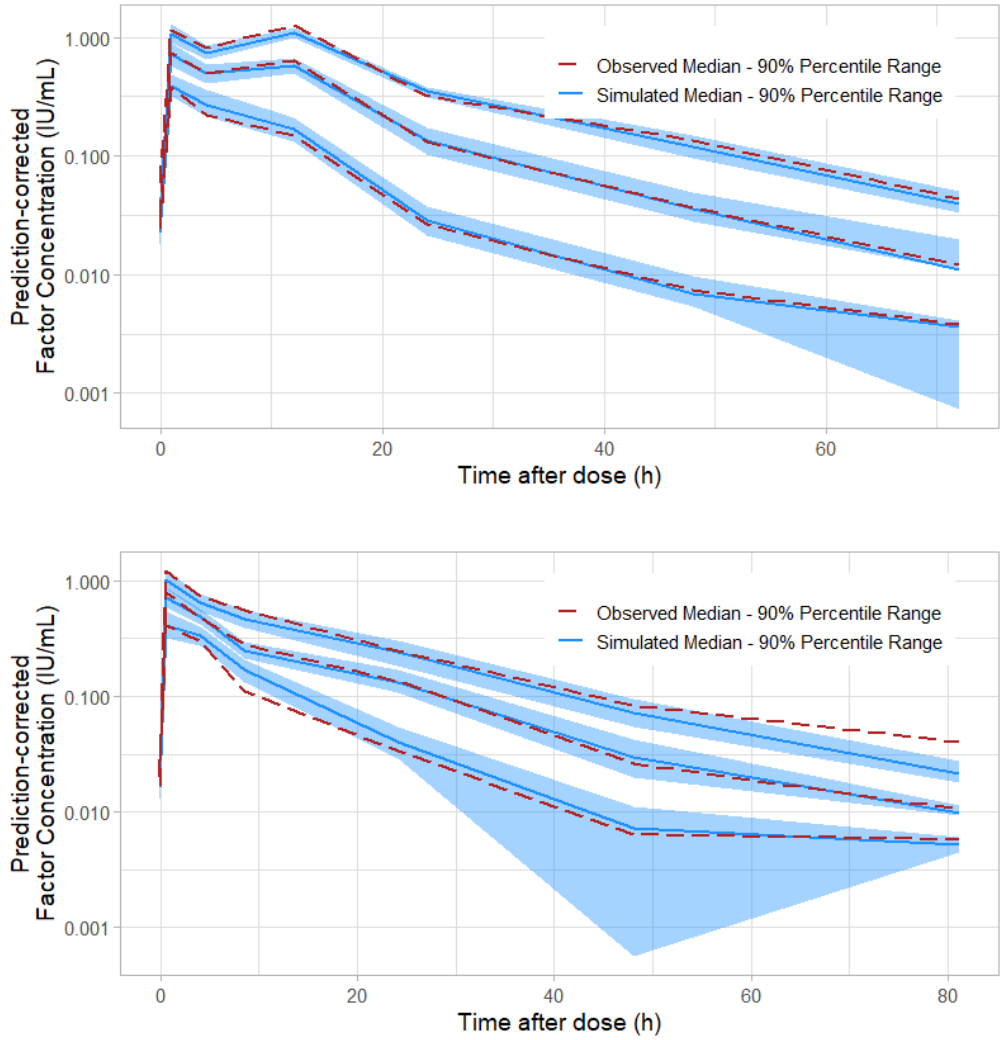



*Figure S19: Observed vs simulated FVIII activities from each PopPK model Bayesian estimations.*

Figures S2 and S3 show the pcVPC graphs that were used to evaluate the models for SHL recombinant, SHL recombinant BDD, SHL plasma derived and Adynovate, Eloctate, and Jivi products, respectively. In all cases, the red dashed lines representing the median, lower and upper percentiles of the distribution of the observed samples agree reasonably well with the corresponding simulation-based inter-percentile bands, constituting a favorable outcome for the pertaining models.

Figure S4 illustrates the differences in estimated PK between age groups and factor concentrates by plotting median and 90% percentile range of simulated PK profiles within these groups.



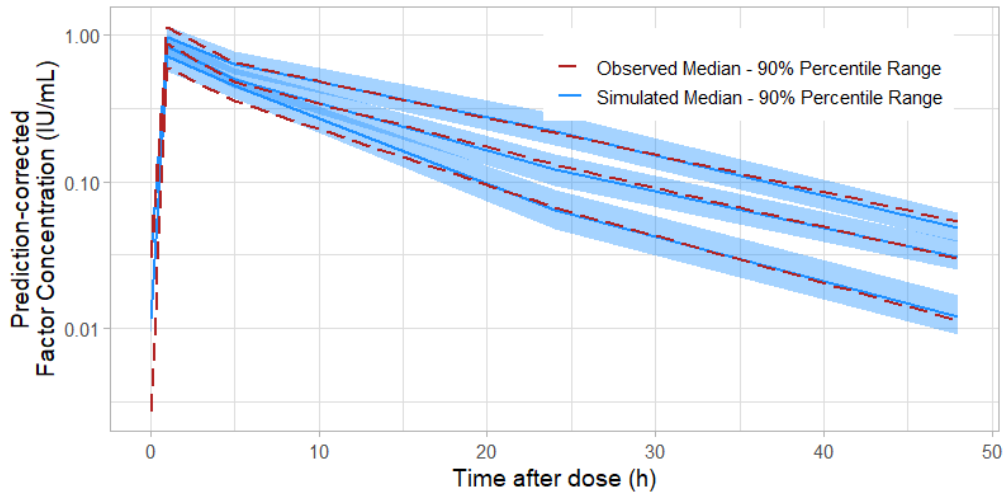
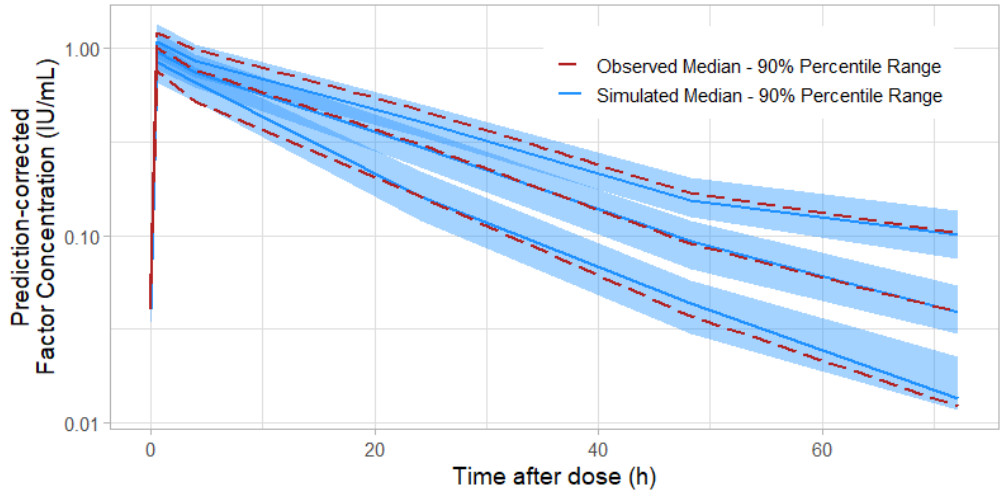
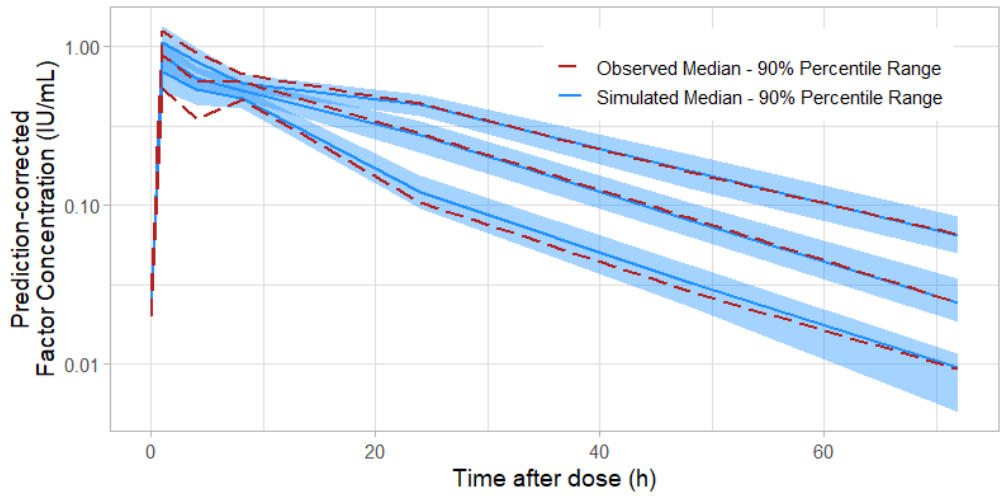



*Figure S20: pcVPC of Bayesian estimations from the PopPK models used for the SHL concentrates (recombinant, recombinant BDD and plasma derived respectively).*



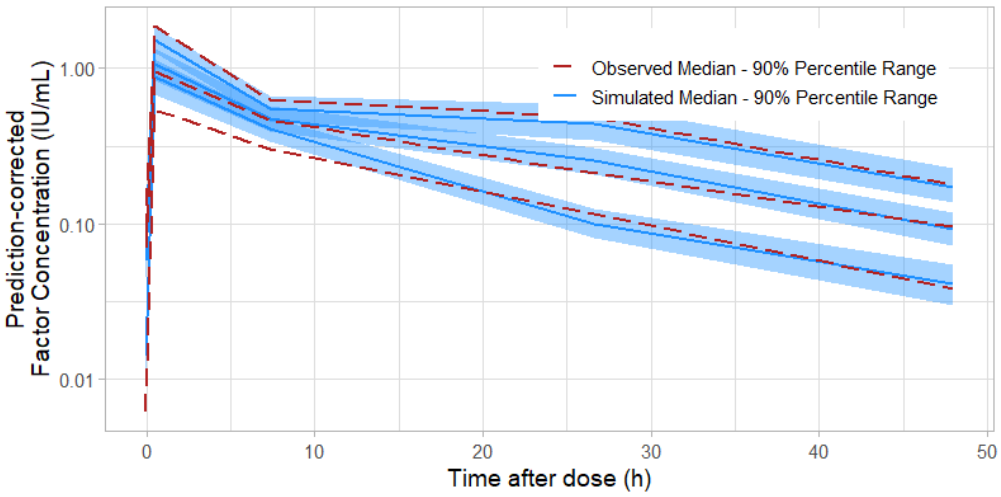



*Figure S21: pcVPC of Bayesian estimations from the PopPK models used for the EHL concentrates (Adynovate, Eloctate and Jivi respectively).*

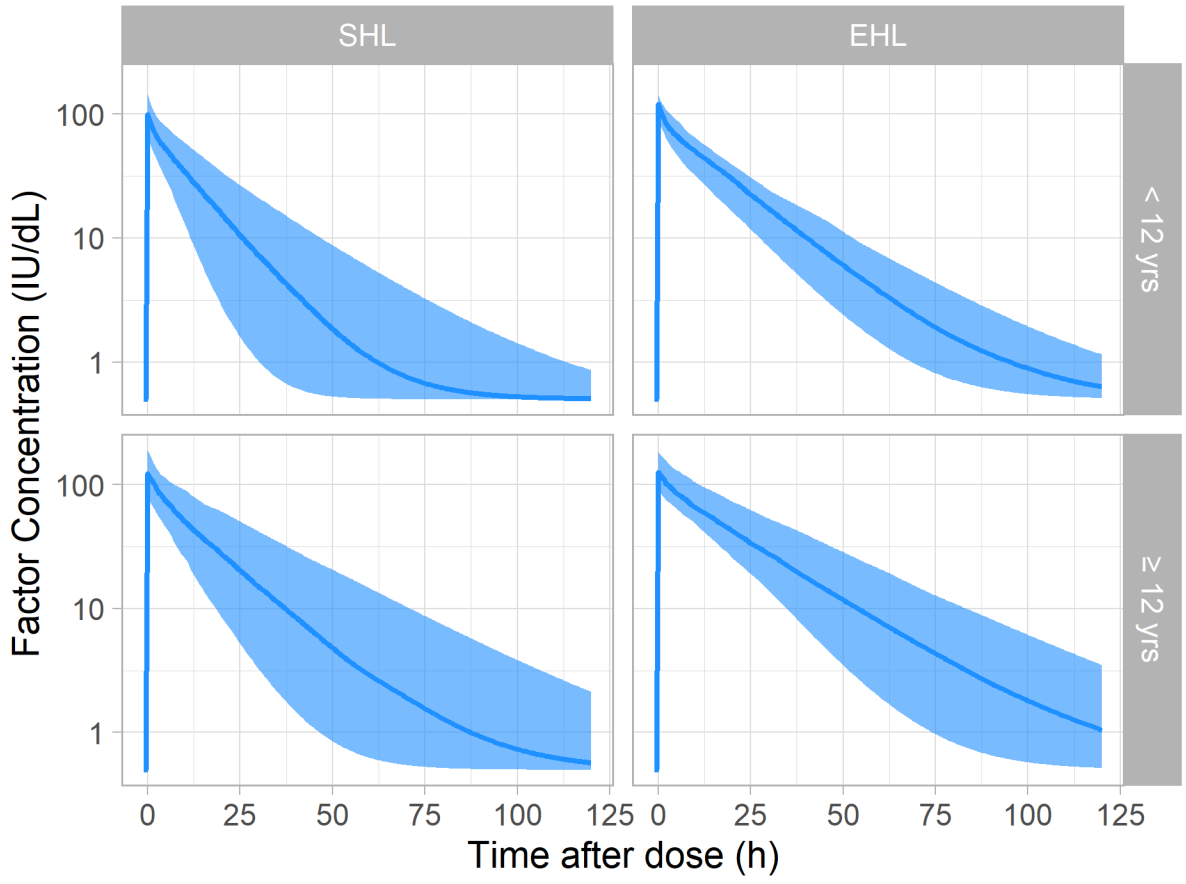



*Figure S22: Simulated FVIII time profile following a single dose of 50 IU/kg of FVIII. Line corresponds to median profile while shaded area is delimited by 5th and 95th percentiles.*

## PD analysis

### Analysis of FVIII at time of bleed event

Overall, 864 bleeds (30.2 %) happened when FVIII was below 1% and 1295 bleeds (45.2 %) happened when FVIII was between 1 and 10%. These ratios were relatively similar between spontaneous and trauma bleeds. From trauma, 451 bleeds (31.5 %) happened when FVIII was below 1% and 652 bleeds (45.5 %) happened when FVIII was between 1 and 10%. While for spontaneous, 413 bleeds (28.9 %) happened when FVIII was below 1% and 643 bleeds (45.0 %) happened when FVIII was between 1 and 10%.

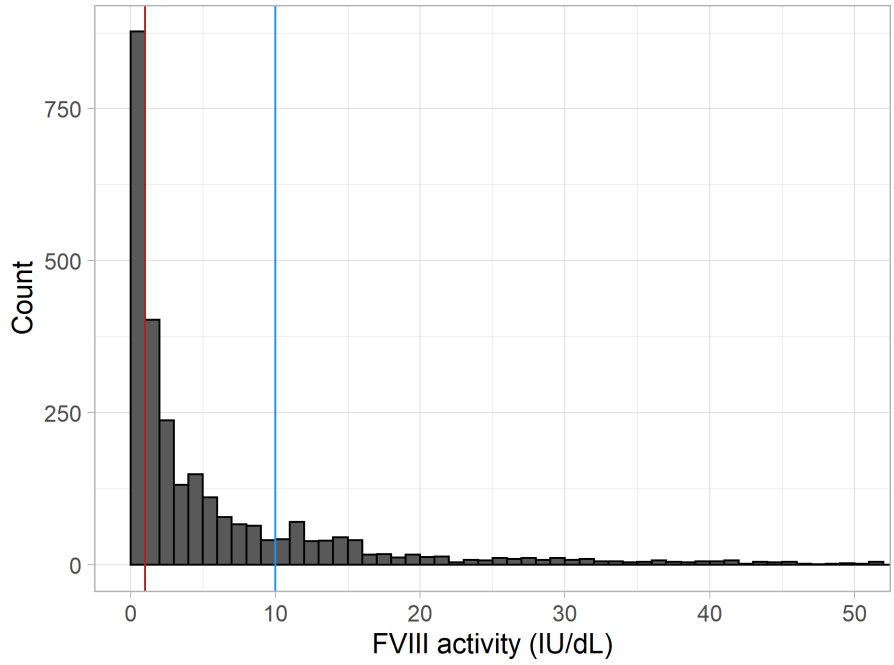



*Figure S23: Distribution of FVIII activity at time of bleed event. Vertical red and blue lines highlight values of 1 and 10 IU/dL.*

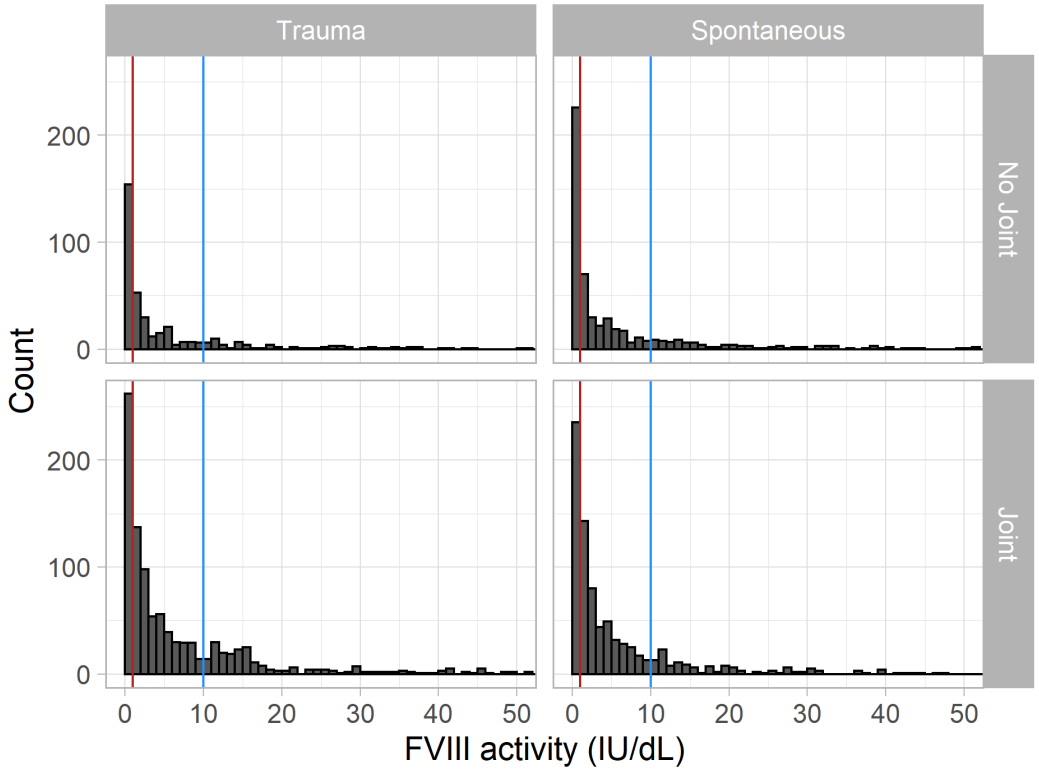



*Figure S24: Distribution of FVIII activity at time of bleed event split by bleed cause and location. Vertical red and blue lines highlight values of 1 and 10 IU/dL.*

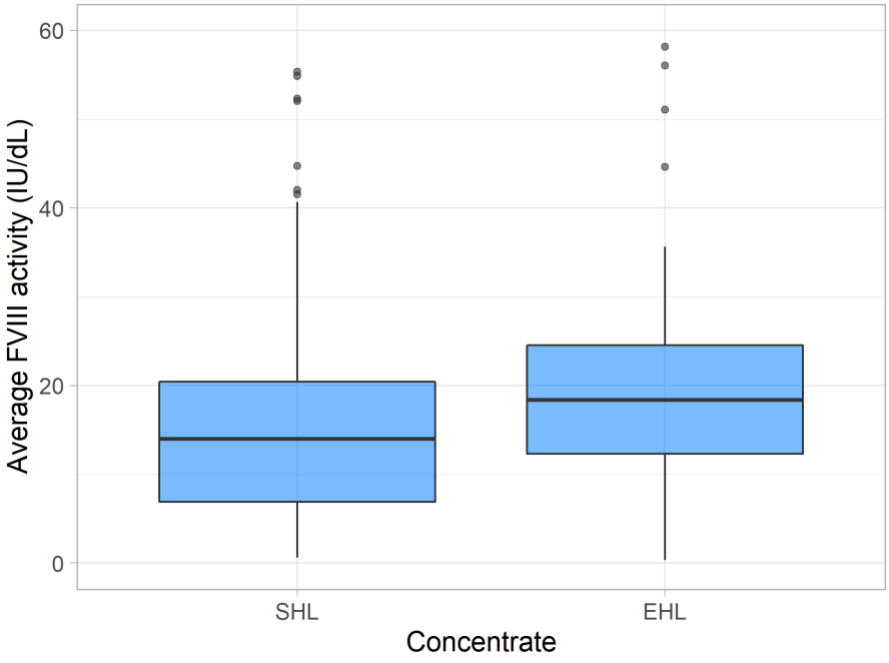



*Figure S25: Distribution of average FVIII activity simulated during the observation periods split by factor concentrate.*

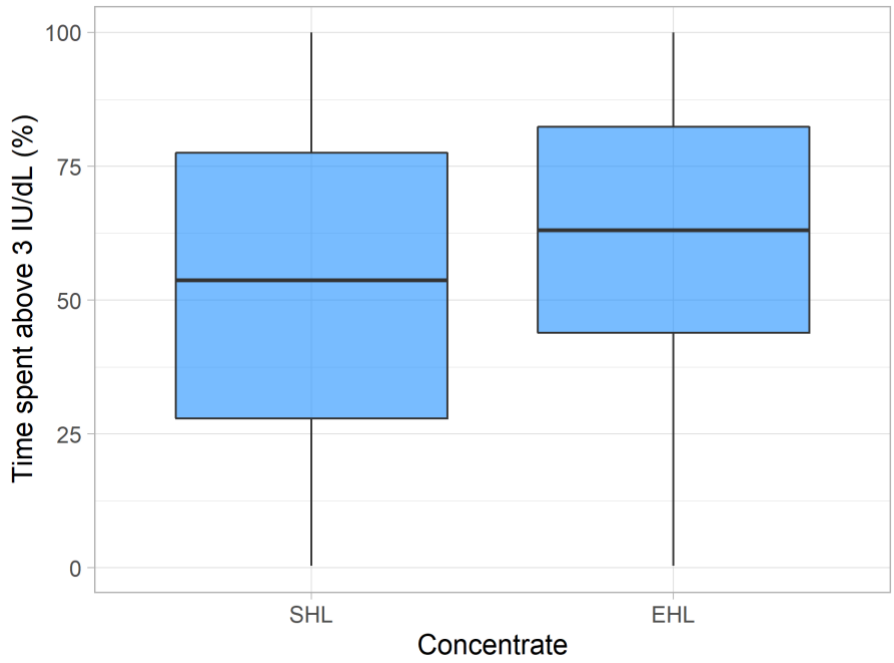



*Figure S26: Distribution of time spent above 3 IU/dL during the observation periods split by factor concentrate.*

**Analysis of survival**

Survival plots using the Kaplan Meier estimator supported the previous results from the analysis of FVIII activities at time of bleed event. Briefly, survival curves were similar when split between bleed causes or between bleed locations (Figures S9 to S13).

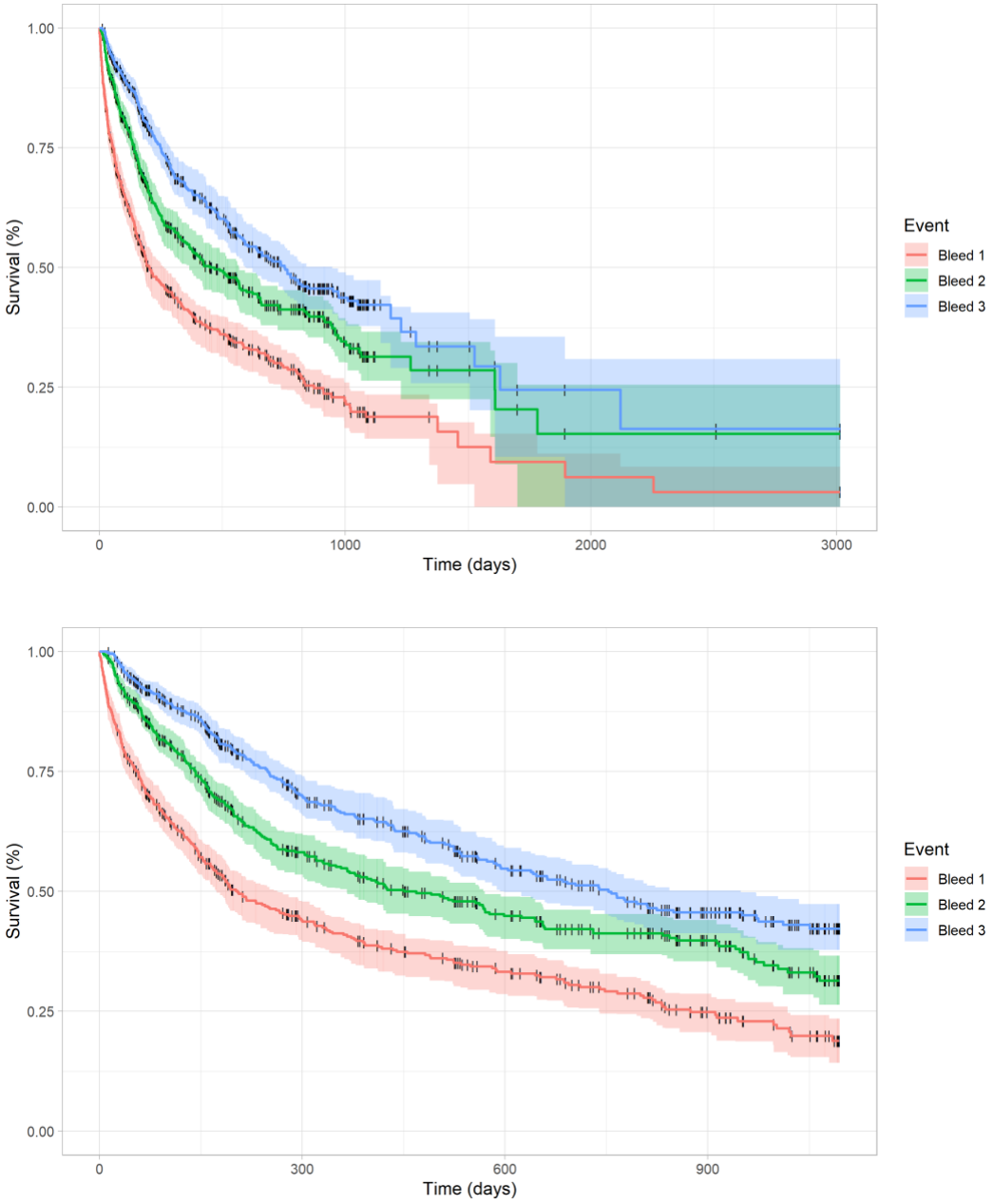



*Figure S27: Kaplan Meier survival for occurrence of first, second and third bleed during the entire observation period (top) and zoomed for the first three years (bottom).*

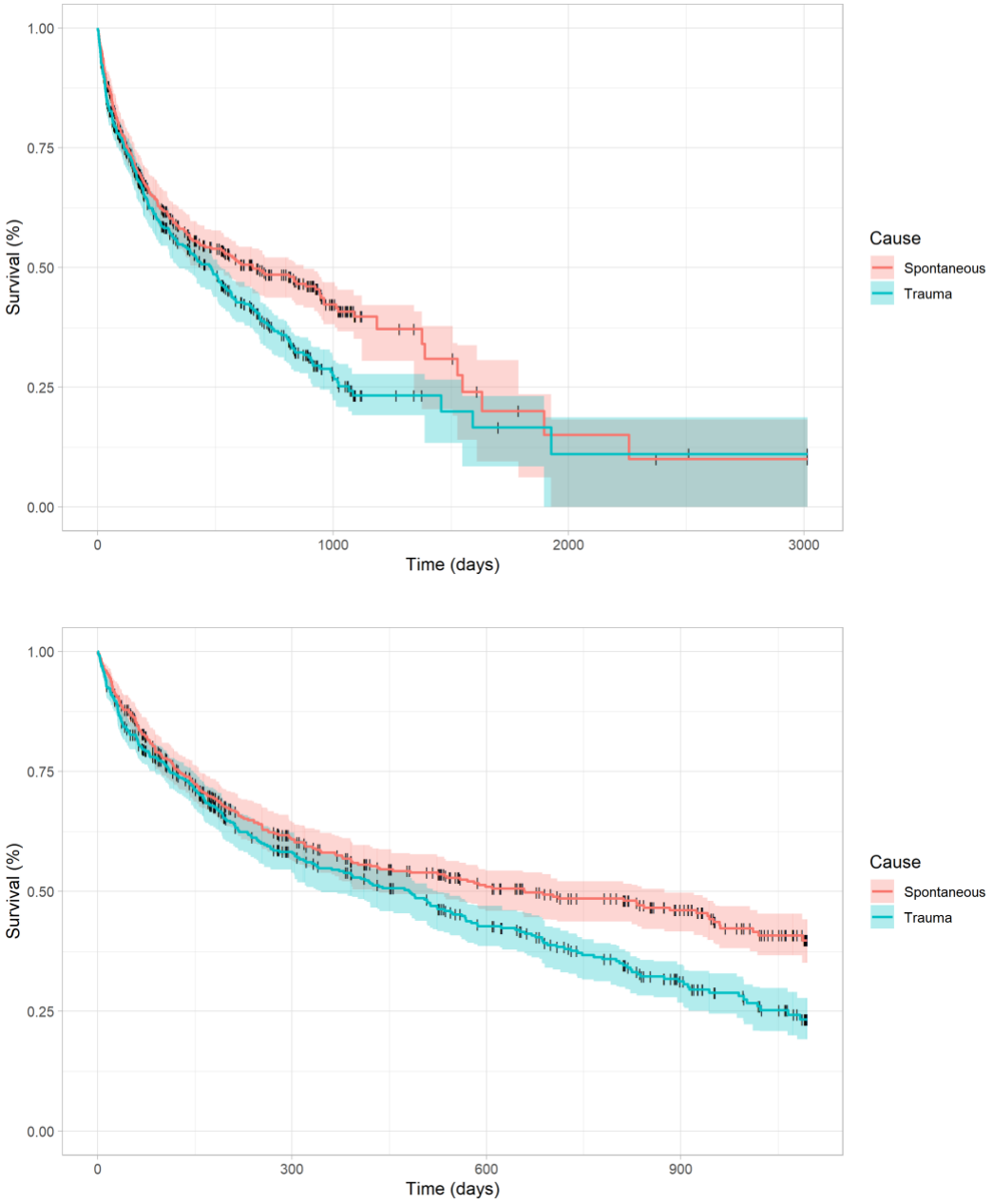



*Figure S28: Comparison of Kaplan Meier survival curves between occurrence of first spontaneous vs trauma bleed during the entire observation period (top) and zoomed for the first three years (bottom).*

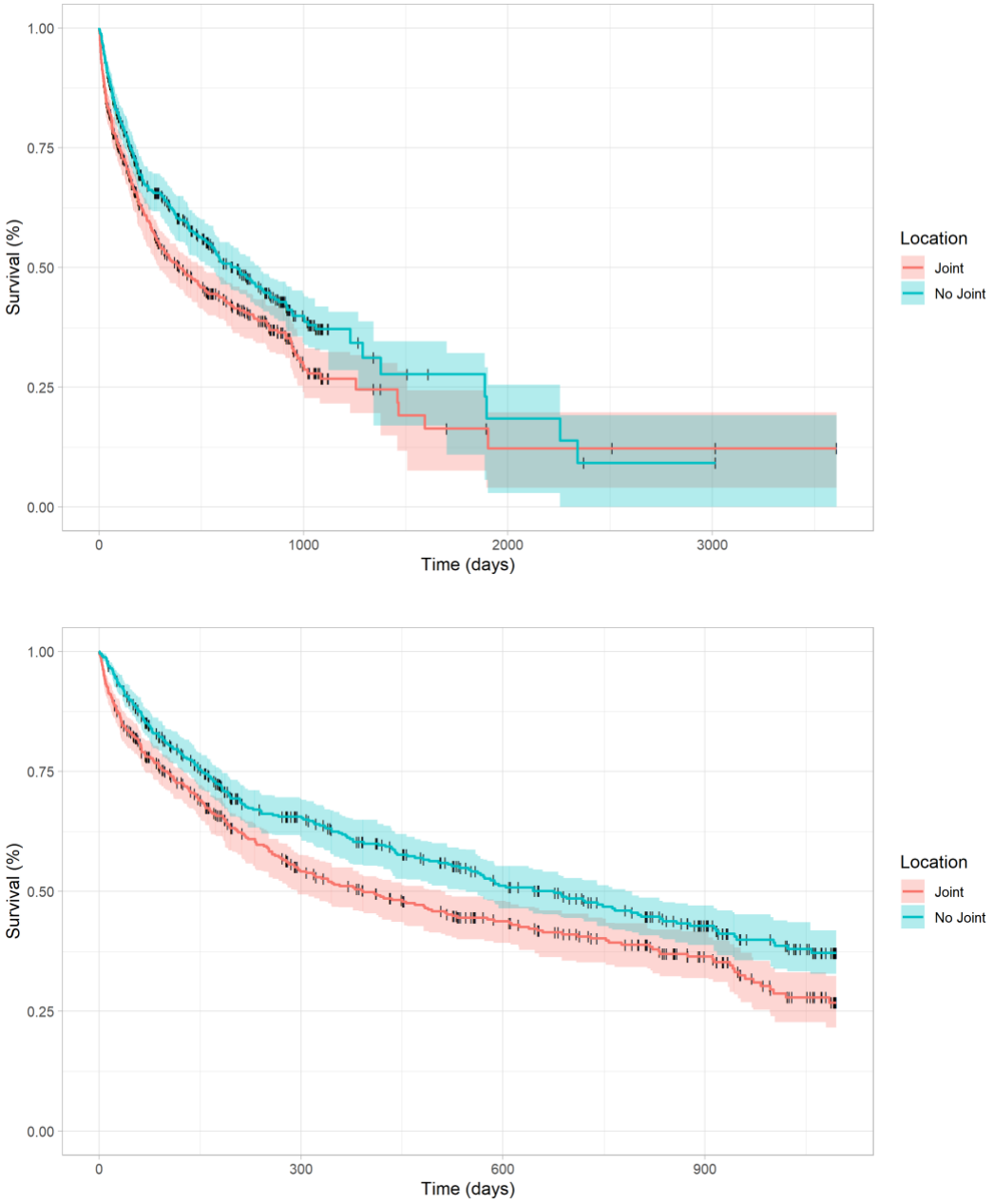



*Figure S29: Comparison of Kaplan Meier survival curves between occurrence of first joint vs no joint bleed during the entire observation period (top) and zoomed for the first three years (bottom).*

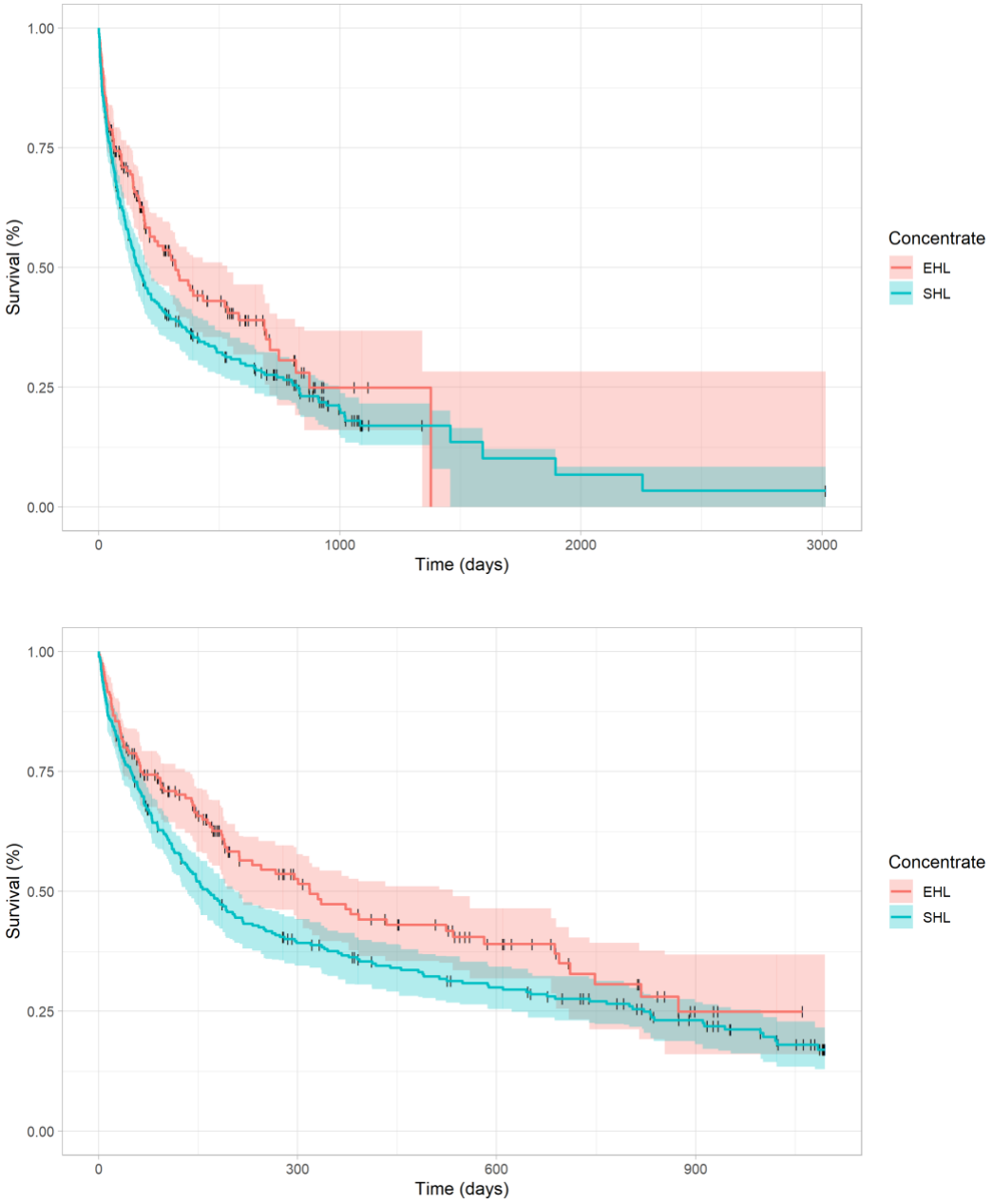



*Figure S30: Comparison of Kaplan Meier survival curves of occurrence of first bleed between SHL and EHL usage during the entire observation period (top) and zoomed for the first three years (bottom).*

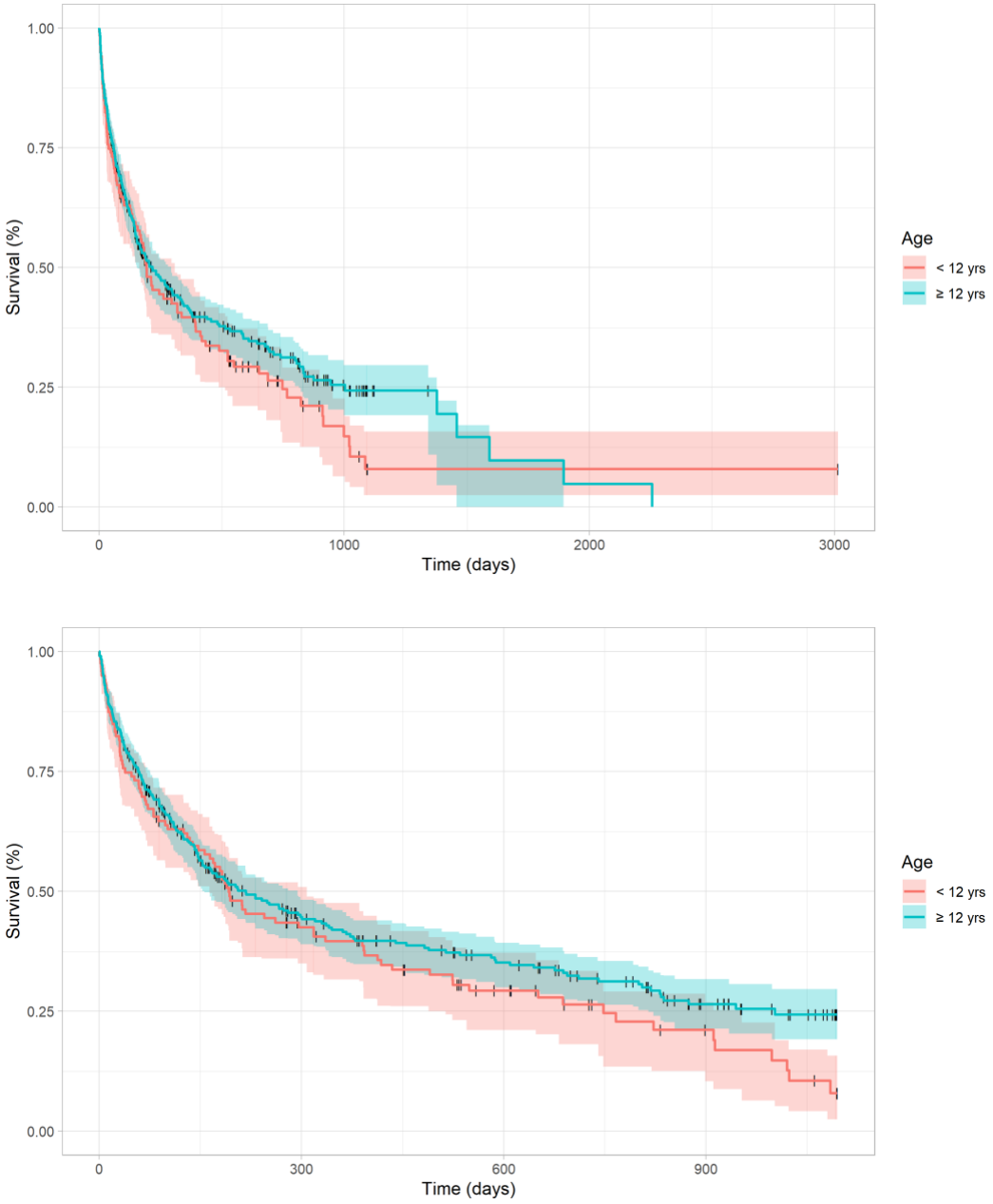



*Figure S31: Comparison of Kaplan Meier survival curves of occurrence of first bleed between children and adolescents/adults during the entire observation period (top) and zoomed for the first three years (bottom).*

**Repeated time to event population modeling**

### *Model development*

The Hill specification showed a significantly lower OFV compared to no and power specification (dOFV < -200) and was kept in subsequently tested models. Weibull function defining the time effect on hazard had the lowest OFV, AIC and BIC. The significance of including probabilities of bleed categories in the model structure was assessed and led to significant drops < -750 in OFV, AIC and BIC. Weibull function as time effect was still the most significant specification regardless after inclusion of bleed cause or location with the model. Additionally, η-shrinkage for models including bleed categories was lower than the 50 % cut-off defined for non significant BSV parameters. Briefly, BSV on hazard showed a shrinkage lower 20 %, while shrinkage for BSV on either bleed location and bleed cause was between 25 and 42 %. The best structural model included a Hill model for the FVIII effect and a Weibull specification to describe the time effect on hazard. As the evaluations of this model were satisfactory (Kaplan Meier VPC and observed vs predicted bleed count), this structural was selected for covariate analysis.

Correlation plots and boxplots of BSV distributions for hazard, bleed cause probability and bleed location probability did not show any apparent trend, although joint bleed probability shows a weak correlation with Age. Shrinkage was also relatively high for all the BSV parameters (16.2, 41.4 and 40.1% for estimated baseline hazard, trauma and joint probabilities, respectively), consequently all covariates were first tested for each BSV parameter. The corresponding results showed significance of concentrate class to describe trauma bleed probability (dOFV=-250, $p=1.3*10^{-56}$) and significance of Age to describe joint bleed probability (dOFV=-189, $p=2.6*10^{-43}$). Using backward elimination of age and EHL also

results in a large OFV increase hence confirming that these 2 covariates are statistically significant (p-value < 0.001) and kept in the final model.

### *Model overview*

The final population RTTE model equations are presented below and the values of its parameters in Table S1. Evaluations of the final model and illustration of the effects of factor activity, observation time and identified covariates are presented in Figures S14 to S19. The final model was:

$$h(t) = h_0(t+1)^{-\beta}\left(1 - \frac{FVIII}{FVIII + EC50}\right)e^{\eta_h}$$

where

- *h(t)* is the hazard at observation time *t* in days
- $h_0$ corresponds to base hazard in bleeds/yr.
- The coefficient $\beta$ is the Weibull parameter corresponding to the time effect (Figures S14 and S15).
- *EC50* is the Hill coefficient for FVIII effect corresponding to a decrease of hazard by 2-fold when *FVIII* activity is equal to *EC50* (Figures S14 and S15).
- BSV is described by the deviation $\eta_h$ that follows a normal distribution of standard deviation equal to 78.7% (Figure S17).

The probability for a bleeding event to have been caused by a trauma $(P_{trauma})$ was described by a logit distribution with a significant covariate effect of concentrate class (Figure S16).

$$P_{trauma} = \frac{e^{tr_0 + \theta_{EHL} * EHL + \eta_{tr}}}{1 + e^{tr_0 + \theta_{EHL} * EHL + \eta_{tr}}}$$

where

- $tr_0$ is the average deviation of the probability
- $EHL$ is the concentrate class. Its value is 1 for EHL concentrates and 0 for SHL concentrates
- $\theta_{EHL}$ is the concentrate class effect
- BSV is described by the deviation $\eta_{tr}$ that follows a normal distribution of standard deviation equal to 58.7% (Figure S18).

The median probability for a bleeding event to be trauma-induced was 50.0% vs 61.9% for SHL vs EHL usage. Inversely, the median probability for a bleeding event to be spontaneous was 50.0% for SHL products vs 38.1% for EHL products. However, since the BSV for bleed cause was high (CV: 58.7%), this covariate effect may not be predictive or clinically significant.

The probability for a bleeding event to happen at a joint location ($P_{joint}$) was described by a logit distribution with a significant covariate effect of age (Figure S16).

$$P_{joint} = \frac{e^{j_0+\theta_{Age}*\log\left(\frac{Age}{19.8}\right)+\eta_j}}{1+e^{j_0+\theta_{Age}*\log\left(\frac{Age}{19.8}\right)+\eta_j}}$$

where

- $j_0$ is the average deviation of the probability
- $\theta_{Age}$ is the age effect

BSV is described by the deviation $\eta_j$ that follows a normal distribution of standard deviation equal to 99.6% (Figure S19).

The relationship between FVIII activity and hazard was modeled using a Hill equation (Figures S14 and S15) describing a maximum hazard when FVIII activity is low and minimum hazard when FVIII activity is high. When factor VIII activity reaches EC50 (14.6 IU/dL) the maximum hazard is divided by 2-fold.

The relationship between observation time and hazard was modeled using a Weibull equation with a negative coefficient (Figures S14 and S15). Thus, hazard decreases rapidly for short observation times – for instance, it decreases by 2-fold after 34 days - and more slowly for longer observation periods – taking 3.26 years to decrease by another 2-fold.

The covariate effects of concentrate class on bleed cause and age on bleed location are illustrated on Figure S16 using box and range plots showing $5^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $95^{th}$ population percentiles of estimated BSV.

Figures S17 to S19 represents the histograms of individual deviations ($\eta$) modeling BSV respectively for base hazard, bleed cause and location probabilities. The histograms are associated with the estimated normal distribution from the model (red line) and histograms of individual base hazard, bleed cause and location probabilities (histograms on the right side).

Table S1: Population RTTE model parameters

| Parameters | Estimate [Shrinkage %] | RSE (%) | 95 % Confidence Interval |
|---|---|---|---|
| Structural Model | | | |
| $\boldsymbol{h_0}$ (bleeds/yr) | 6.100 | 19.5 % | [3.77; 8.43] |
| EC50 (IU/dL) | 14.6 | 4.75 % | [13.2; 16.0] |
| $\boldsymbol{\beta}$ (1/yr) | -0.195 | 3.08 % | [-0.207; -0.183] |
| Median trauma bleed probability $\boldsymbol{P_{tr0}}$ (%) (For SHL usage) | 50.0 % | - | - |
| Median joint bleed probability $\boldsymbol{P_{j0}}$ (%) (At median Age 19.8 yrs) | 58.9 % | 1.57 % | [58.6; 59.2] |
| Covariate Model | | | |
| EHL on trauma bleed probability | 0.485 | 6.31 % | [0.425; 0.545] |
| Age on joint bleed probability | 0.194 | 4.07 % | [0.179; 0.209] |
| BSV | | | |
| Hazard (CV%) | 78.7 % [0.88 %] | 0.192 % | [78.6; 78.9] |
| Trauma (CV%) | 58.7 % [48.6 %] | 0.243 % | [58.5; 58.8] |
| Joint (CV%) | 99.6 % [46.1 %] | 0.545 % | [99.1; 100.0] |

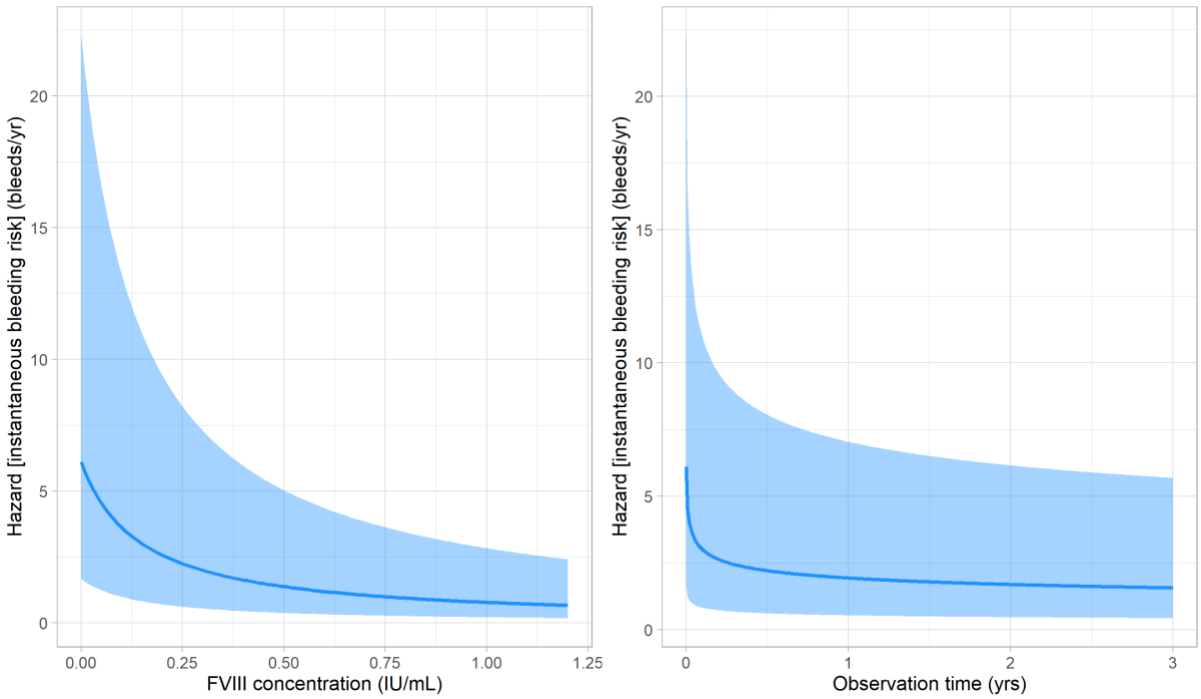



*Figure S32: Left: Hazard [instantaneous bleeding risk] with BSV as a function FVIII without time effect (time = 0 day); Right: Hazard with BSV as a function time without FVIII effect (FVIII = 0 IU/mL)*

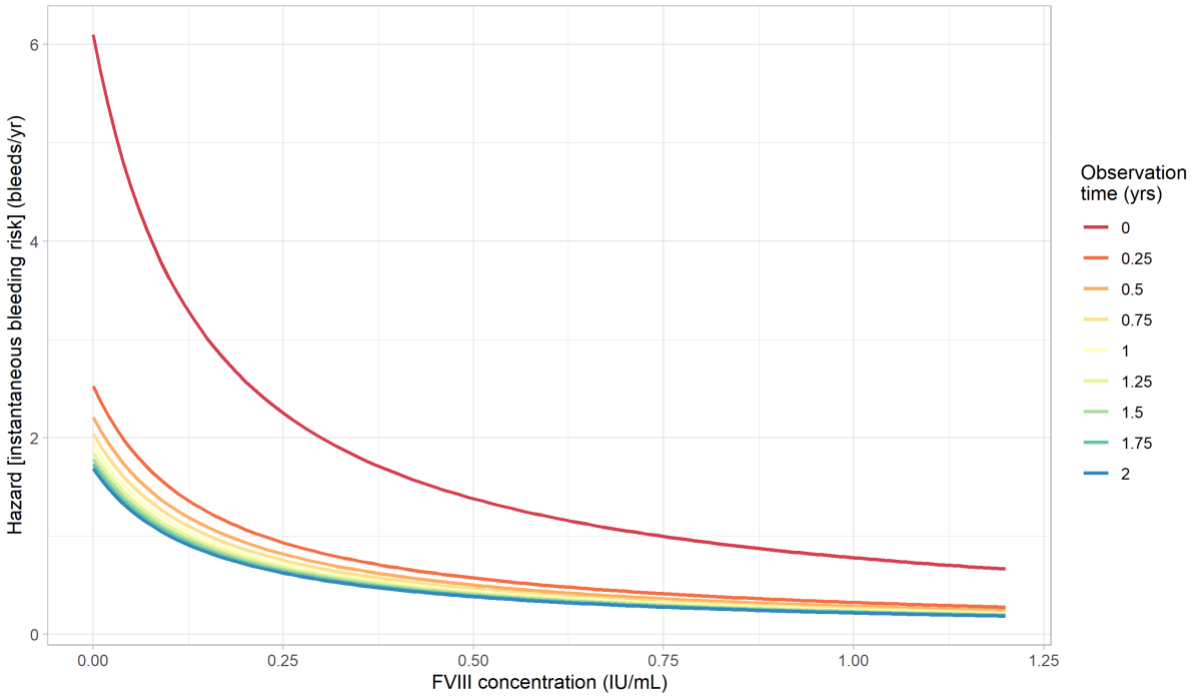


*Figure S33: Median hazard [instantaneous bleeding risk] as a function FVIII at different observation times*

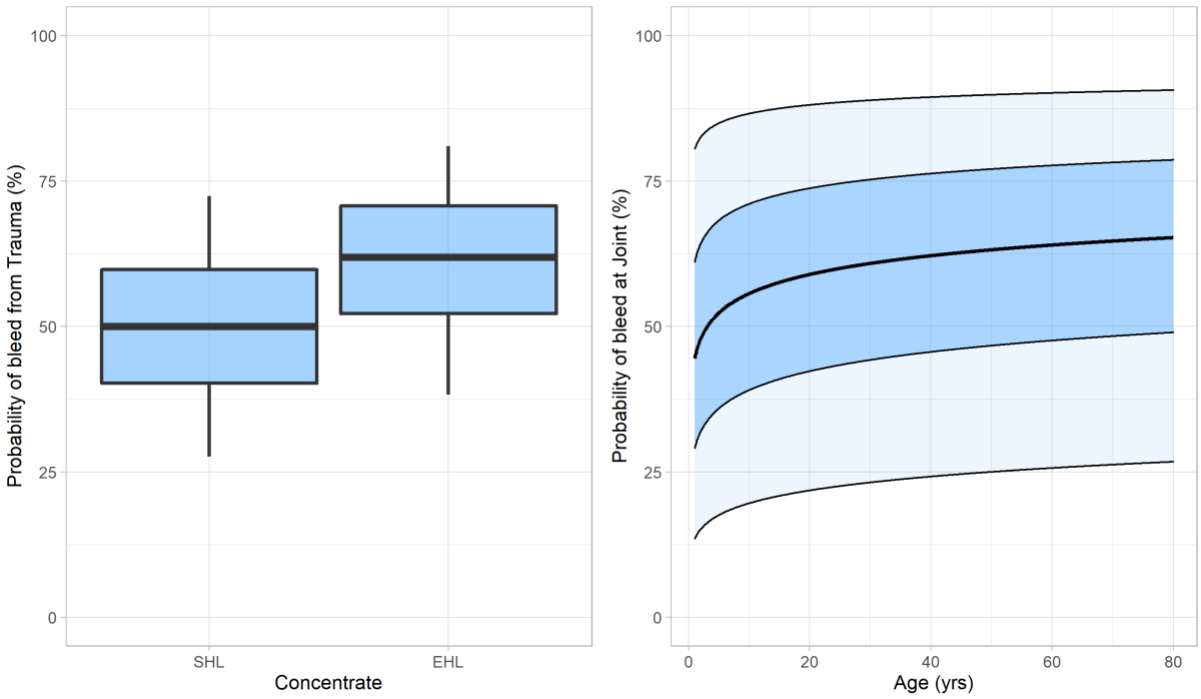



*Figure S34: Left: Distribution of trauma bleed probability as a function of concentrate class; Right: Distribution of joint bleed probability as a function of Age.*

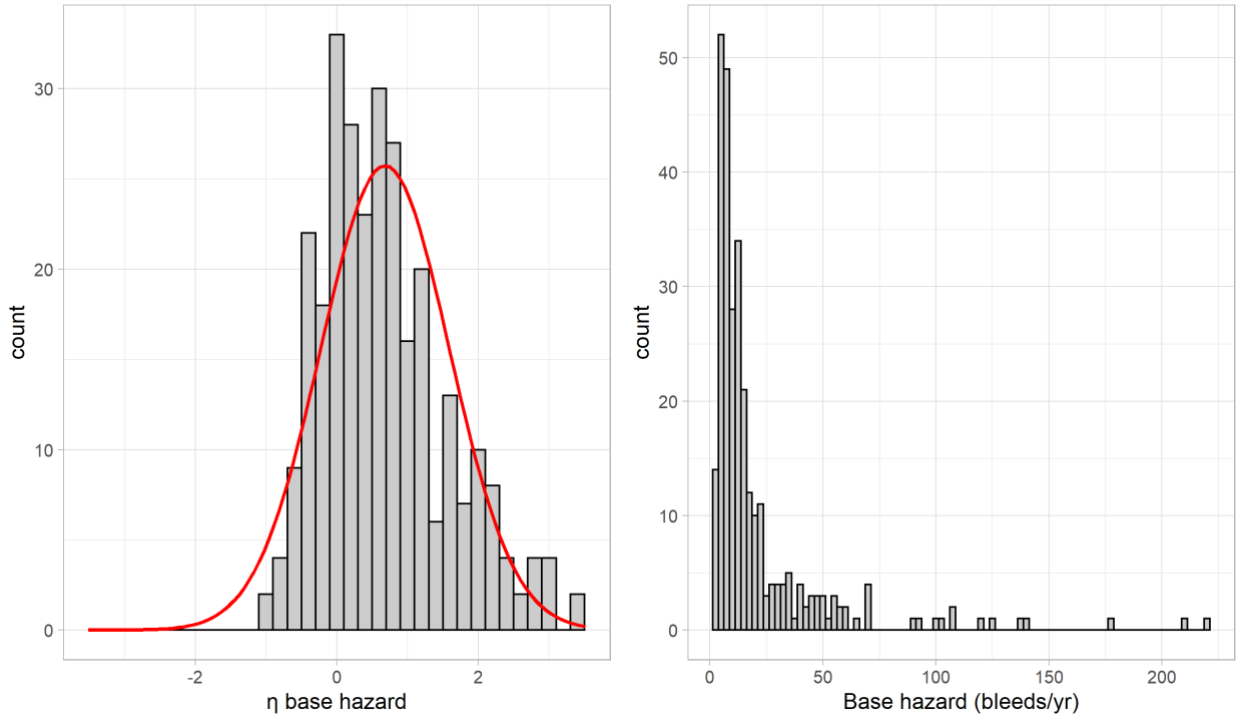


*Figure S35: Left: Distribution of hazard BSV parameter ($\eta_h$). Right: Distribution of base hazard BSV*

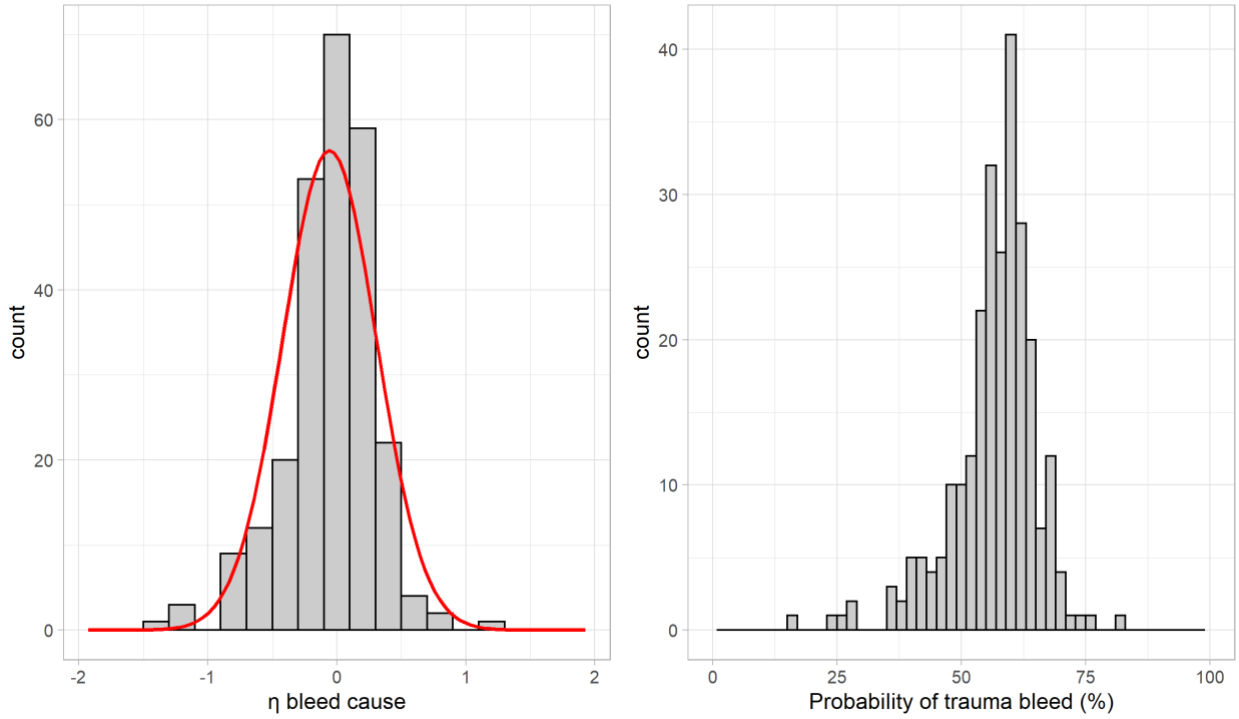



*Figure S36: Left: Distribution of bleed cause probability BSV parameter ($\eta_{tr}$). Right: Distribution of probability of bleed cause from trauma*

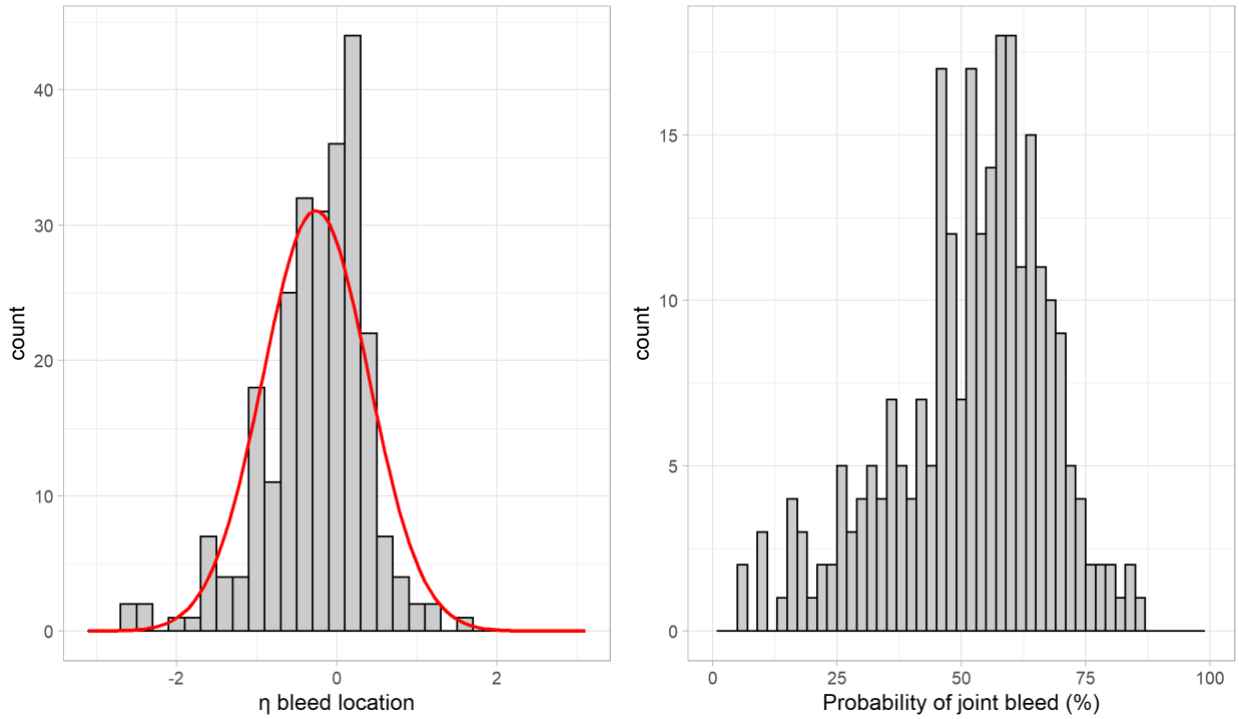


*Figure S37: Left: Distribution of bleed location probability BSV parameter ($\eta_j$). Right: Distribution of probability of bleed location at joint*

### *Model evaluations*

Evaluations of the quality of the fit of the final model were performed. Predicted vs observed overall annual bleed rate was well fitted with a coefficient of determination of 0.76 (Figure S20). Likewise, cumulative hazard representing the predicted bleed count at times of bleed event fitted well ($R^2$=0.95) the observed bleed count at the same times of event regardless of bleed cause or category (Figure S21). For patients who did not record any bleed during their observation period (n=137), the distribution of cumulative hazard was represented on Figure S22. For such patients, the predicted number of bleeds was usually low with 81% of predicted bleed count lower than 2 bleeds.

Stratified VPCs of the Kaplan Meier estimator were performed to assess if the model was able to capture the observed survival and its variability. Overall, the survival for the occurrence of the first 4 bleeds was well predicted with observed survival mostly within the 90% CI of the simulated values (Figure S23 and S24). Survival for first bleed event which was slightly under-predicted while survival for more than 5 bleeds was slightly over-predicted. Focusing on the subjects that recorded one bleed or less (n=196), in average 51% (±16%) of these subjects were simulated with one bleed or less out of the 500 simulations.

Assessing for occurrence of the first bleed event stratified by bleed cause, the fit of the final model improved compared to the structural model and captured most of the observed survival within its 90% CI of the simulated values (Figure S25). The model slightly under-predicted each stratified event as it was overall under-predicting the occurrence of first bleed.

Assessing for occurrence of the first bleed event stratified by bleed location, the fit of the final model also captured most of the observed survival within its 90% CI of the simulated values (Figure S26). The model slightly under-predicted each stratified event as it was overall under-predicting the occurrence of first bleed.

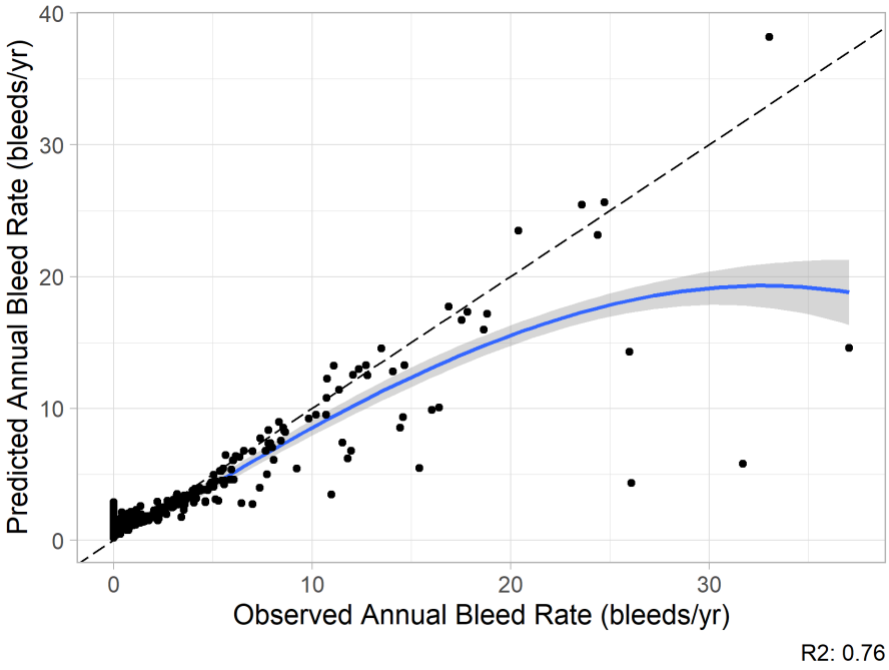



*Figure S38: Observed vs predicted annual bleed rate. Blue line corresponds to loess regression.*

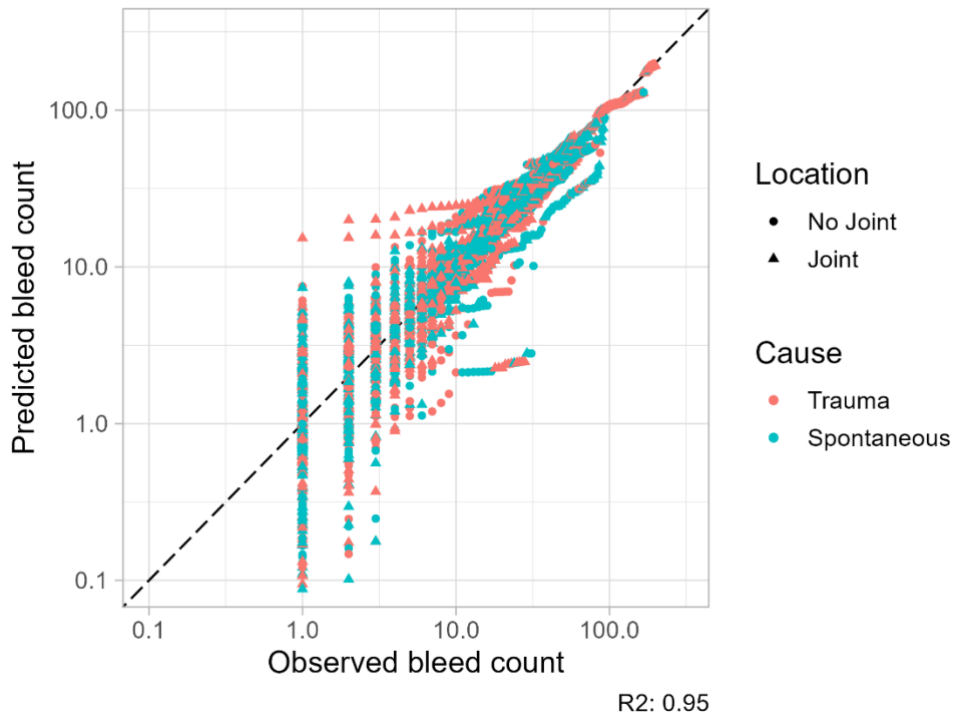



*Figure S39: Observed vs predicted (cumulative hazard) bleed count at times of bleed events.*

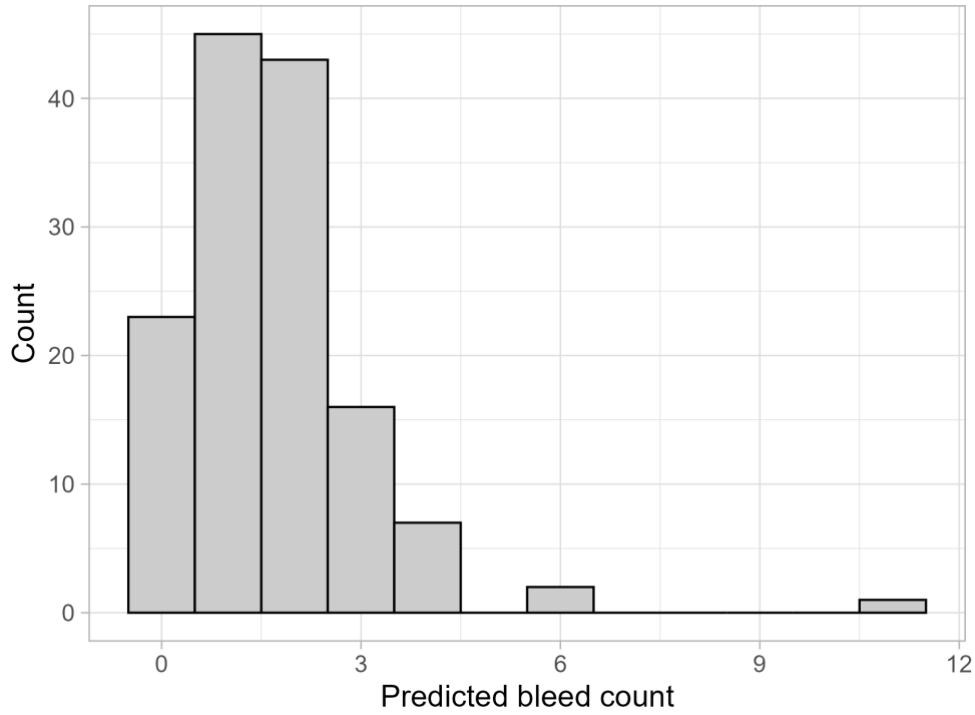



*Figure S40: Distribution of predicted bleed count (cumulative hazard) at the end of observation period for patients who did not bleed.*

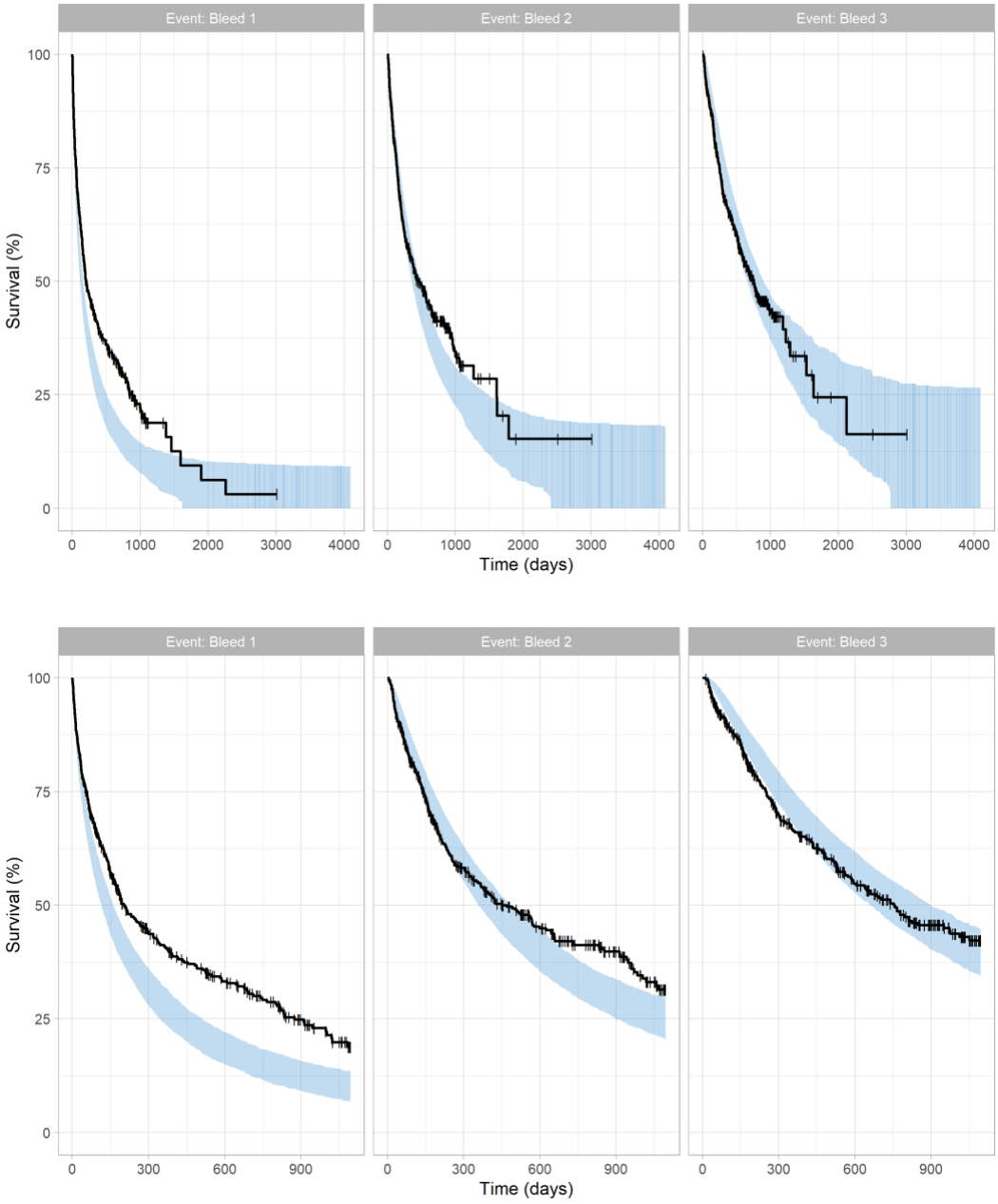



*Figure S41: Visual predictive check of Kaplan Meier estimator for occurrence of the first three bleeds for the final model. Bottom plots corresponds to the first three years of the observation period.*

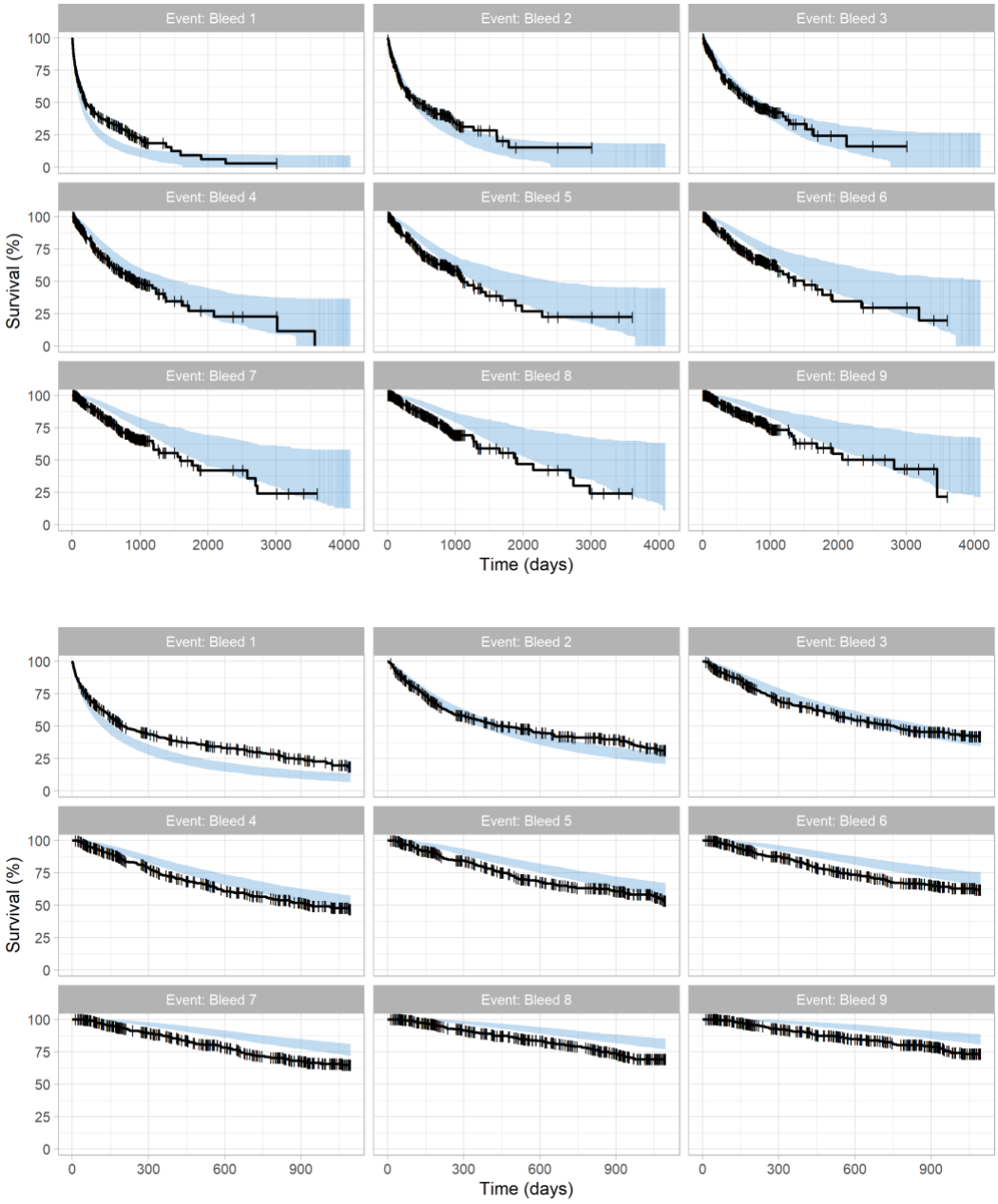



*Figure S42: Visual predictive check of Kaplan Meier estimator for occurrence of the first nine bleeds for the final model. Bottom plots corresponds to the first three years of the observation period.*

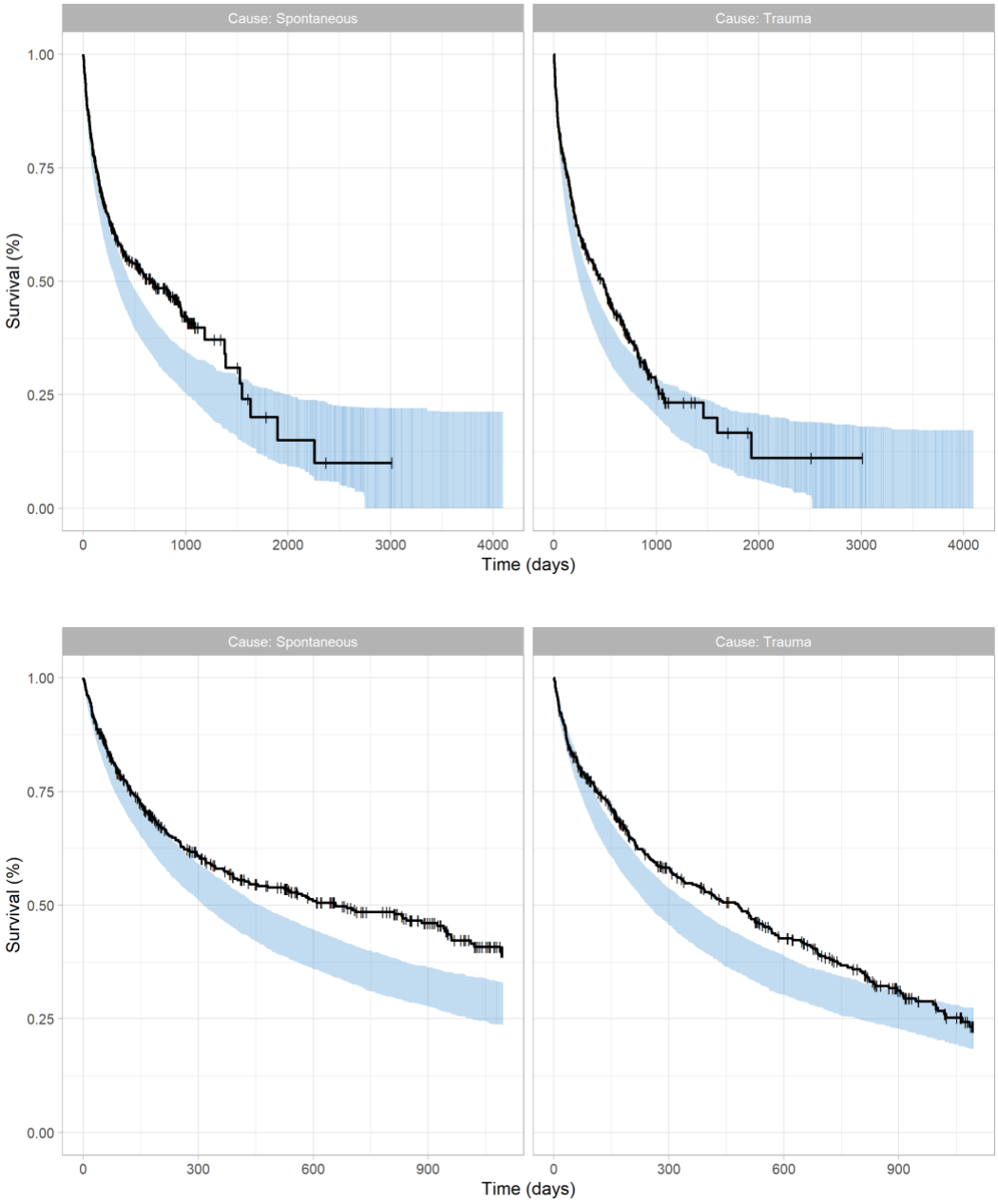



*Figure S43: Visual predictive check of Kaplan Meier estimator for occurrence of first trauma vs spontaneous bleed. Bottom plots corresponds to the first three years of the observation period.*

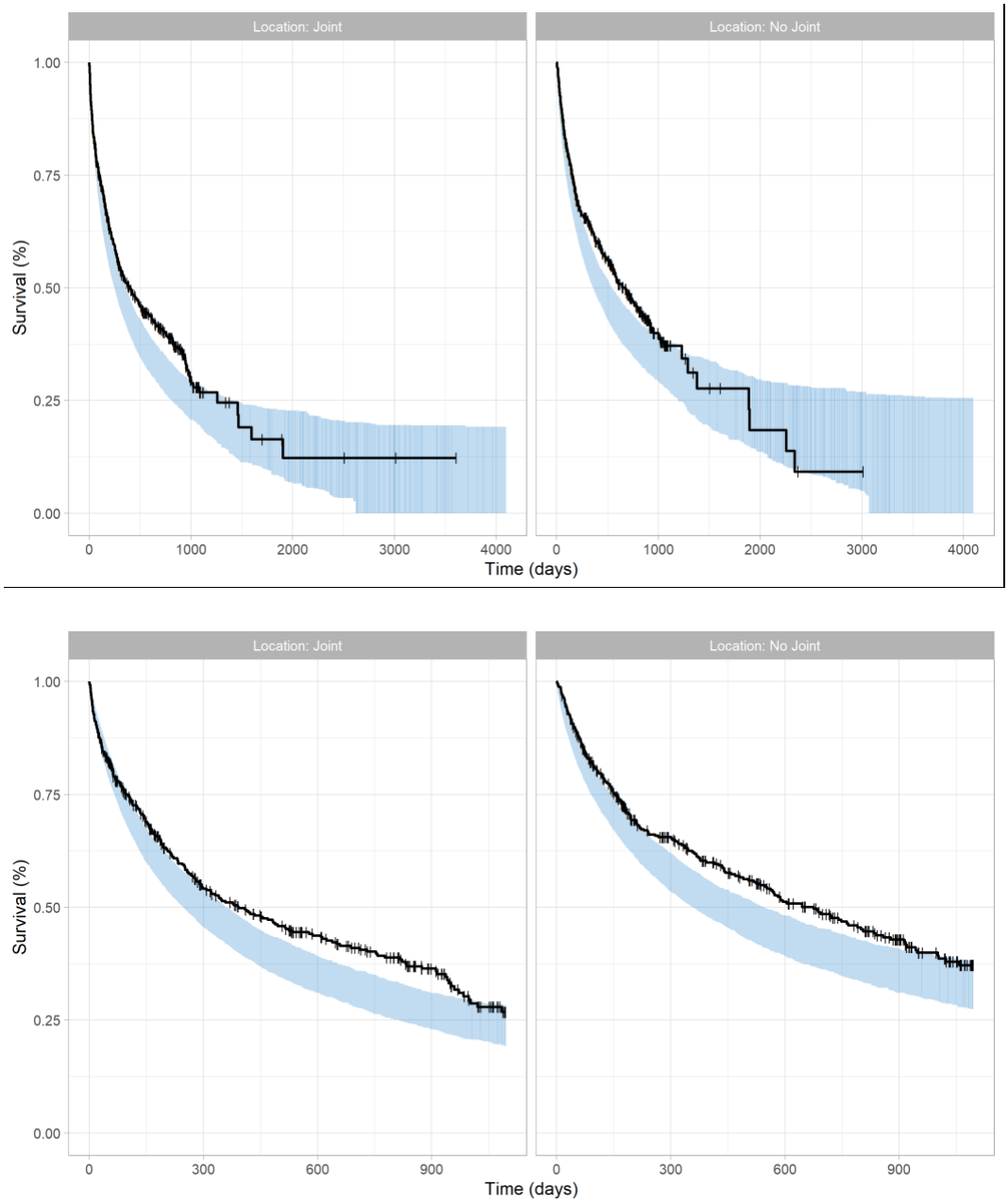



*Figure S44: Visual predictive check of Kaplan Meier estimator for occurrence of first joint vs no joint bleed. Bottom plots correspond to the first three years of the observation period.*

# Supplementary material – Chapter 5

## Title: Study protocol: comparing regression to random forests for predicting the risk for bleeding in people living with hemophilia.

## Background and rationale:

Hemophilia is an inherited X-linked bleeding disorder. Hemophilia A is characterized by a deficiency of clotting factor VIII, while factor IX is deficient in hemophilia B. The cornerstone of hemophilia care is the treatment with the deficient clotting factor. The severity of bleeding episodes is usually associated with clotting factor levels. Hemophilia with factor levels <0.01 IU/ml is classified as severe. Severe hemophilia is associated with spontaneous blees into joints or muscles, even in the absence of identifiable hemostatic challenges.(1) In untreated patients with severe hemophilia, recurrent bleeds in joints progressively cause disabling arthritis. On the other hand, these spontaneous bleeds seldom occur in patients with factor levels >0.01 IU/ml.(2) These observations led to the introduction of prophylactic replacement of clotting factor, with the aim of keeping the patients' factor levels at least >0.01 IU/ml.(3) Prophylactic treatment is usually dosed by weight, but this can translate in either underdosing or overdosing, due to variability in the individual patients' pharmacokinetics (PK) profiles.(4) The consequential variation in the patients' factor levels affects their risk for bleeding.(5,6) A systematic review conducted by our group (still unpublished) showed that other risk factors for bleeding in people living with hemophilia (PWH) are physical activity levels,(5) age,(7) bleeding history,(8) joint status,(9) and obesity.(9) Patients, physicians, and policymakers might benefit from knowing the risk of

bleeding of individual patients. From the patients' perspective, we think it would be helpful to know what's their risk of bleeding and how this might change modifying potential risk factors. For example, knowing that changing the treatment adherence from 70 to 90% would reduce the annualized risk of bleeding by a certain amount, might help a patient in improving his adherence. Moreover, knowing the punctual risk of bleeding based on risk factors including modifiable variables such as the plasma level concentration and the type of physical activity to be performed, would allow the patients to change their risk, modifying the factor plasma levels (with an extra infusion) before a high-risk activity, or avoiding high-risk activities when the factor levels are too low. On the other hand, when the risk for bleeding is very low, a patient might reduce the factor usage, and this would allow saving resources. From the physicians' perspective, a risk assessment model could be used for educational purposes as described above, and to select the best treatment for a specific patient, for example reserving more intensive treatment regimens to patients at high risk of bleeding. From a policy-maker perspective, the identification of different risk categories would allow them to decide how to allocate resources. This is particularly important now that new therapies are about to enter the market. Emicizumab is a humanized antibody that mimics the function of factor VIII and presents potential clinical advantages as compared to factor concentrates, being administered s.c. instead of e.v., and less frequently.(10,11) Another option will also soon be available: gene therapy for patients with hemophilia A and B have been tested in phase 1 and 2 trials with promising results,(12,13) several other phase 2 studies are ongoing, and some companies already moved to phase 3 (NCT03392974, NCT03370913). It is reasonable to assume that these therapies will be very expensive. Policymakers will have to decide how many resources to allocate to them and which groups

of patients will be eligible for those treatments. The identification of patients at high risk of bleeding that can't lower their risk changing some of the risk factors might be a way to select the patients that can benefit more from the new treatments (even if, ideally, this should then be proven in further clinical studies). Our systematic review did not identify any existing risk assessment model (RAM) for the prediction of the risk for bleeding in PWH.

**Objective**

The objective of the present study is to compare regression techniques to random forests for the prediction of the risk for bleeding in PWH.

## Methods:

The present protocol has been realized referring to the SPIRIT statement,(14) adapted according to the RECORD statements(15) and the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research.(16) Even though these are meant mainly for study reporting (not for protocol preparation), they appeared to be appropriate to adapt the SPIRIT statement, which is made for protocols, but suits mainly to randomized control trials.

**Study Design and data sources**

We will perform a national, multi-center, retrospective, cohort study. We will use data from the Web-Accessible Population Pharmacokinetic Service - Hemophilia (WAPPS-Hemo) and the Canadian Bleeding Disorders Registry (CBDR). WAPPS-Hemo is a population-based Bayesian calculator that provides caregivers with individual patients' PK estimates for many factor concentrates.(17) The service is hosted at McMaster, is industry independent, and freely available at the website https://www.wapps-hemo.org. An example of a PK estimate

is provided in **Figure 45**. CBDR contains clinical and administrative data on Canadian PWH and is also hosted at McMaster. CBDR has a module for patients, called MyCBDR, that allows them to input treatment and bleed logs using a mobile application or a website. WAPPS-Hemo and CBDR are linked on the back end.

**Participants**

**Inclusion criteria:** People living with severe congenital Hemophilia A or B, treated with factor concentrate prophylaxis (regardless of the drug used), taking bleeding and treatment diaries (either electronic or paper-based).

**Exclusion criteria:** other concomitant bleeding diatheses (congenital or acquired bleeding disorders), presence of inhibitors (antibodies against factor VIII or IX).

**Outcomes**

The primary study outcome will be the annualized bleeding rate (ABR), calculated as the number of bleeds per patient divided by the follow-up time, expressed in years (bleeds $n*patient^{-1}*y^{-1}$).

The secondary outcome will be the ABR categorized in three levels: 0, 1-2, and ≥3 bleeds/year.

**Predictors**

Knowing the **PK profile** (from WAPPS-Hemo) and the infusion dose and time (from CBDR) of each PWH allows estimating their factor levels over time. This information will be used to calculate the proportion of time spent with factor levels ≤1 IU/ml, ≤3 IU/ml, ≤5 IU/ml, ≤10 IU/ml, and >10 IU/ml. PK parameters will be extracted from WAPPS-Hemo, including volume of distribution, clearance, area under the curve (AUC), and half-life. The following variables

will be extracted from CBDR. **Baseline:** age, sex at birth, hemophilia type (A or B) and severity (baseline factor VIII or IX levels, expressed as IU/mL), history of inhibitors (YES/NO), blood group (A, B, AB, 0; Rh + or -) type of mutation (as classified by the European Association for Hemophilia and Associated Disorders - EAHAD - Coagulation Factor Variant Databases);(18) **Physical examination:** weight, height, body mass index, joints health status expressed as hemophilia joint health score (HJHS) and number of target joints (joints with 3 or more bleeds over a period of 6 months);(19) **Treatment plan:** factor concentrate name, type (standard vs extended half-life), dose (IU), and frequency of administration ($days^{-1}$); **Treatment history:** registered factor usage for prophylaxis (total dose as IU), treatment of bleeds (total dose as IU) and invasive procedures (total dose as IU); amount of factor concentrates ordered (total dose as IU), treatment adherence, calculated as the ratio between the amount of factor used according to the treatment diaries and the amount of factor concentrate prescribed by the physician based on the treatment plan; **Surgery:** number of invasive procedures, type, and date of execution; **Bleeding history** (ABR in the previous year).

**Timeline**

The study timeline is reported in **Table 21**. Baseline characteristics will be extracted at $T_0$ or the closest time point available before $T_0$. Subjects with severe hemophilia are usually assessed in the clinic at least every year, and this assessment will be used as $T_0$.

Predictors that need to be calculated over time, like factor levels, bleeding history, and treatment adherence, will be measured in the 6-12 months before $T_0$, with 12 months being the target, and any period between 6 and 12 months being considered acceptable.

The outcome (bleeding events) will be measured for a 6-12 months period after $T_0$, with 12 months being the target, and any period between 6 and 12 months being considered acceptable.

**Study size**

We aim at including all the eligible patients in CBDR with a PK profile available in WAPPS-Hemo. Therefore, we did not conduct a formal sample size calculation. 711 patients living with severe hemophilia (597 A + 114 B) are registered in CBDR. Approximately 75% of them (i.e., 533) should be treated with prophylaxis on a regular basis. Among these patients, a pop-PK estimation has already been performed in ~60%, leading to 320 patients.

**Statistical methods**

*Data segmentation*

The dataset will be split into a training set (outcome data up to June 2020) and a test set (outcome data for the period July 2020 – June 2021). The test set will be used for what in healthcare research is called internal validation, with the performance of the predictive models being tested on unseen data.

*Variable selection*

We will feed the model with all the available variables identified through our systematic review, plus variables considered important predictors by two hemophilia treaters. The remaining variables will be further selected based on the correlation between them: we will create a Pearson correlation matrix between all pairs of variables, and remove one predictor from any pairs with a correlation coefficient >0.9.(20,21) This will allow us to make the model easier to interpret and less computationally intense. Being the number of available variables reasonably limited, we will not implement further selection steps.

### *Descriptive analysis*

Baseline patients' characteristics and outcome measures will be described with standard descriptive statistics. The following measures of central tendency and distribution will be used as appropriate to describe continuous variables: median and standard deviation for normally distributed data and median and interquartile range for skewed data. Discrete data will be presented as absolute numbers and percentages.

### *Predictive models*

**Regression:** The primary outcome will be analyzed using a multivariate multilevel mixed-effects negative binomial regression with a random intercept, considering PKs nested within participants (when more than one will be available), and participants nested within treatment centers.(22)

The secondary outcome will be analyzed using a multinomial logistic regression.(23)

**Random forests:** a random forest is an ensemble of classification trees. Each decision tree is trained with a bootstrapped sample of the dataset.(24) Not only the participants but also the variables fed to each three are randomly sub-sampled. Each tree is grown without pruning (i.e., not limiting the number of branches).(24) The random forest provides a meta-estimation that aggregates many decision trees.(25) This estimate outperforms the individual classification trees. The fact that each tree uses a subsample of the participants reduces the chances of overfitting, and subsampling the variables avoids relying too heavily on a single predictor and addresses potential collinearity. We expect a class imbalance, with 20-30% of patients experiencing 0 bleeds. This is likely to translate into an imbalanced prediction error. To mitigate this effect, we will assign different weights to the classes. We will start with weights inversely proportional to the class size and adjust the weights

iteratively until a balance is reached.(24) The numbers of predictors per tree and the minimum size of each node will be tuned. We will set the number of trees in the forest to a computationally feasible number (~1500), without tuning.(21,26)

### *Missing data*

Missing data will be assumed to occur at random. For the regression models, we will perform multiple imputations using the chained equations method.(27) For the random forests, we will input data using the average (for continuous variables) or the most frequent (for categorical variables) of the non-missing values weighted by the proximities to the missing case. The fills will be recalculated and replaced through five iterations.(24)

### *Variable importance*

Ranking the variables for their importance and quantifying their contribution to the prediction process helps with interpreting the model. This is especially important for "black box" models like random forests, where there is no straightforward interpretation of each predictor's role in the model functioning. For the regression models, the variables will be ranked based on the regression coefficients. For the random forests, the importance will be computed averaging the gini decreases for each variable across all the trees.(24)

### *Performance measures*

The models' performance will be evaluated on the test set. For the primary outcome, measures of performance will be the root mean square error, the mean absolute error, and the bias in the predicted number of bleeds.(28) For the secondary outcome, the performance measure will be the AUC, calculated as the weighted average of the AUC of the model for each level against the other two levels, with the weight being the proportion of each level.(21,29)

## Ethics

### *Research ethics approval*

The CBDR and WAPPS-Hemo projects are already approved by the Hamilton Integrated Research Ethics Board (HiREB). The present protocol will also be submitted for approval to HiREB.

### *Consent*

Considering the retrospective nature of the project and the fact that the CBDR and WAPPS-Hemo users consent to the use of their data for research purposes, we will ask for a waiver of informed consent.

### *Confidentiality*

Data will be analyzed at McMaster. The database will be anonymized before the analysis. The study results will be published presenting aggregated data.

## Data sharing policy

Data will be made available to researchers and regulators upon reasonable request.

## Team composition

I am a medical doctor with an interest in health research methodology and biostatistics. For this project, I will coordinate a team composed of two experts in hemophilia (one also experienced in prognosis methods) and at least one each of the following: a biostatistician (better if competent in machine learning), an IT and a data analyst routinely working on the CBDR and WAPPS-Hemo database, and a data/computer scientist with experience in machine learning.

## Strengths and limitations

Our study has several strengths. It will be the first study using estimates of the factor levels based on individual PKs and treatment logs on a large scale, using real-world data, and incorporating this information into a prediction model. The only previous experience using a similar predictor is a study conducted on 104 patients from interventional studies.(5) This study was not oriented at producing a prediction model but had more of an explanatory aim, with the objective of exploring the risk for bleeding associated with physical activity in PWH after controlling for their factor levels. The multicenter, national design will support the generalizability of the results. The use of random forests has the potential to improve the predictive performance of the model as compared to classical statistical methods. Admittedly, our study also presents some limitations. First, as of now, we don't have available data on the physical activity levels, which might be an important predictor for the bleeding risk. We are working on a project for the passive data collection on physical activity using wearable devices (e.g., smart watches) in PWH. This will address this limitation and make the use of machine learning techniques even more important, as the data to be processed will increase exponentially. Some patients will be represented both in the training and test sets. This might expose the model to "validation leakage", where information from the training set propagates to the test set.(16) If this will happen, the model performance will be overoptimistic, limiting generalizability. One way of addressing this limitation might be to repeat the experiment changing the split, e.g. using a random split of the dataset or a geographical split based on centers, using some for training and some for testing. The con of

this strategy is that the sample size will be reduced, and we will implement it only if the model performance in the test set will be too close to the one on the training set.

## Dissemination and future directions

The study results will be disseminated through presentations at conferences, publication in a peer-reviewed journal, and hopefully in future guidelines. Future directions will be the external validation of the model, even better if simplified only using important variables. We are working on linking the WAPPS-Hemo service to some bleeding disorders registries around the world, and this will offer the possibility for external validation of the model. Once validated, the model will be ready for use.

## References

1. Srivastava A, Santagostino E, Dougall A, Kitchen S, Sutherland M, Pipe SW, et al. WFH Guidelines for the Management of Hemophilia, 3rd edition. Haemophilia. 2020 Aug 1;26:1–158.

2. Arnold WD, Hilgartner MW. Hemophilic arthropathy. Current concepts of pathogenesis and management. J Bone Joint Surg Am. 1977 Apr;59(3):287–305.

3. Manco-Johnson MJ, Abshire TC, Shapiro AD, Riske B, Hacker MR, Kilcoyne R, et al. Prophylaxis versus episodic treatment to prevent joint disease in boys with severe hemophilia. N Engl J Med. 2007;357(6):535–44.

4. Hazendonk HCAM, Lock J, Mathôt RAA, Meijer K, Peters M, Laros-van Gorkom BAP, et al. Perioperative treatment of hemophilia A patients: blood group O patients are at risk of bleeding complications. J Thromb Haemost. 2016 Mar;14(3):468–78.

5. Broderick CR, Herbert RD, Latimer J, Barnes C, Curtin JA, Mathieu E, et al. Association Between Physical Activity and Risk of Bleeding in Children With Hemophilia. JAMA. 2012 Oct 10;308(14):1452.

6. Collins PW, Blanchette VS, Fischer K, Björkman S, Oh M, Fritsch S, et al. Break-through bleeding in relation to predicted factor VIII levels in patients receiving prophylactic treatment for severe hemophilia A. J Thromb Haemost. 2009 Mar;7(3):413–20.

7. Ross C, Goldenberg NA, Hund D, Manco-Johnson MJ. Athletic participation in severe hemophilia: Bleeding and joint outcomes in children on prophylaxis. Pediatrics. 2009 Nov;124(5):1267–72.

8. Abrantes JA, Solms A, Garmann D, Nielsen EI, Jönsson S, Karlsson MO. Relationship between factor VIII activity, bleeds and individual characteristics in severe hemophilia A patients. Haematologica. 2020 May 1;105(5):1443–53.

9. Gupta S, Siddiqi AEA, Soucie JM, Manco-Johnson M, Kulkarni R, Lane H, et al. The effect of secondary prophylaxis versus episodic treatment on the range of motion of target joints in patients with haemophilia. Br J Haematol. 2013 May;161(3):424–33.

10. Shima M, Hanabusa H, Taki M, Matsushita T, Sato T, Fukutake K, et al. Factor VIII–Mimetic Function of Humanized Bispecific Antibody in Hemophilia A. N Engl J Med. 2016 May 26;374(21):2044–53.

11. Mahlangu J, Oldenburg J, Paz-Priel I, Negrier C, Niggli M, Mancuso ME, et al. Emicizumab Prophylaxis in Patients Who Have Hemophilia A without Inhibitors. N Engl J Med. 2018 Aug 30;379(9):811–22.

12. Rangarajan S, Walsh L, Lester W, Perry D, Madan B, Laffan M, et al. AAV5–Factor VIII Gene Transfer in Severe Hemophilia A. N Engl J Med. 2017 Dec 28;377(26):2519–30.

13. George LA, Sullivan SK, Giermasz A, Rasko JEJ, Samelson-Jones BJ, Ducore J, et al. Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. N Engl J Med. 2017 Dec 7;377(23):2215–27.

14. Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013;346.

15. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Med. 2015 Oct 6;12(10):e1001885.

16. W L, D P, T T, S G, S R, C K, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec 1;18(12).

17. Iorio A, Keepanasseril A, Foster G, Navarro-Ruan T, McEneny-King A, Edginton AN, et al. Development of a Web-Accessible Population Pharmacokinetic Service-Hemophilia (WAPPS-Hemo): Study Protocol. JMIR Res Protoc. 2016 Dec 15;5(4):e239.

18. McVey JH, Rallapalli PM, Kemball-Cook G, Hampshire DJ, Giansily-Blaizot M, Gomez K, et al. The European Association for Haemophilia and Allied Disorders (EAHAD) Coagulation Factor Variant Databases: Important resources for haemostasis clinicians and researchers. Haemophilia. 2020 Mar 1;26(2):306–13.

19. Hilliard P, Funk S, Zourikian N, Bergstrom B-M, Bradley CS, McLimont M, et al. Hemophilia joint health score reliability study. Haemophilia. 2006 Sep;12(5):518–25.

20. Koller D, Sahami M. Toward Optimal Feature Selection. Stanford InfoLab; 1996 Feb.

21. Jones A, Costa AP, Pesevski A, McNicholas PD. Predicting hospital and emergency department utilization among communitydwelling older adults: Statistical and machine learning approaches. PLoS One. 2018 Nov 1;13(11).

22. den Uijl IEM, Fischer K, Van Der Bom JG, Grobbee DE, Rosendaal FR, Plug I. Analysis of low frequency bleeding data: the association of joint bleeds according to baseline FVIII activity levels. Haemophilia. 2011 Jan;17(1):41–4.

23. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Vol. 398. John Wiley & Sons; 2013.

24. Breiman L, Cutler A. Random forests — Classification description: Random forests [Internet]. http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home. htm. 2007 [cited 2021 Aug 12]. p. 1–27. Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

25. Maini V, Sabri S. Machine Learning for Humans. The ultimate guide to machine learning. [Internet]. 2017 [cited 2021 Aug 12]. p. 97. Available from: https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

26. Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random forest. J Mach Learn Res. 2017;18(1):6673–90.

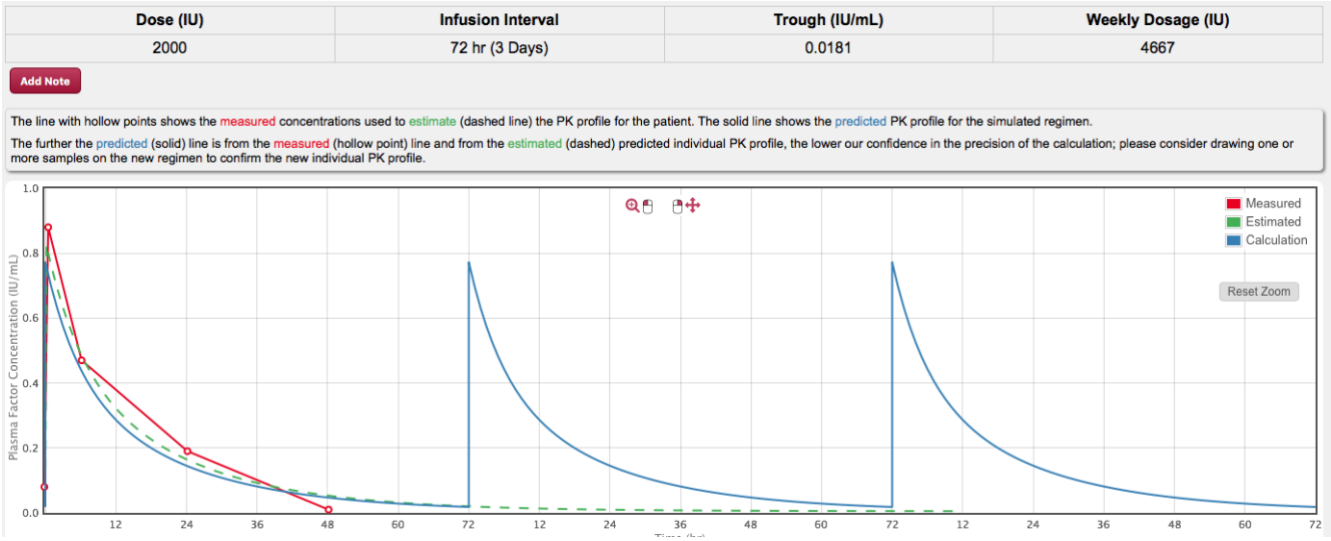
27. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011 Feb 20;30(4):377–99.

28. Holodinsky JK, Yu AYX, Kapral MK, Austin PC. Comparing regression modeling strategies for predicting hometime. BMC Med Res Methodol. 2021 Dec 1;21(1).

29. Calster B Van, Belle V Van, Condous G, Bourne T, Timmerman D, Huffel S Van. Multi-class AUC metrics and weighted alternatives. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008. p. 1390–6.

## Tables and Figures

*Figure 45: example of the graphic results of a PK estimate.*




IU: international units; PK: pharmacokinetics

*Table 21: participants timeline.*

| | STUDY PERIOD | | |
|---|---|---|---|
| | Historical control | At baseline* | Follow-up |
| TIMEPOINT** | $-t_{12\ to\ 6}$ | $t_0$ | $t_0$-$t_{12}$ |
| ASSESSMENTS: | | | |
| Baseline variables | | | |
| Factor levels | | | |
| Treatment history | | | |
| Bleeding history | | | |
| Target joints | | | |
| Surgery | | | |

| | | | |
|---|---|---|---|
| Bleeds (outcome) | | | |

*or the closest time point available before $T_0$