BMJ Open

Predicting hospital admissions from individual patient data (IPD): an applied example to explore key elements driving external validity

Andreas Daniel Meid,[1] Ana Isabel Gonzalez-Gonzalez ,[2,3] Truc Sophia Dinh,[2] Jeanet Blom,[4] Marjan van den Akker ,[2,5] Petra Elders,[6] Ulrich Thiem,[7] Daniela Küllenberg de Gaudry,[8] Karin M A Swart,[6] Henrik Rudolf,[9] Donna Bosch-Lenders,[5] Hans J Trampisch,[9] Joerg J Meerpohl,[8] Ferdinand M Gerlach,[2] Benno Flaig,[2] Ghainsom Kom,[10] Kym I E Snell,[11] Rafael Perera,[12] Walter Emil Haefeli,[1] Paul Glasziou,[13] Christiane Muth [2,14]

To cite: Meid AD, Gonzalez-Gonzalez AI, Dinh TS, *et al*. Predicting hospital admissions from individual patient data (IPD): an applied example to explore key elements driving external validity. *BMJ Open* 2021;**11**:e045572. doi:10.1136/bmjopen-2020-045572

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-045572).

ADM and AIG-G contributed equally.

Received 07 October 2020
Accepted 10 July 2021

For numbered affiliations see end of article.

Correspondence to
Dr Andreas Daniel Meid; andreas.meid@med.uni-heidelberg.de and Dr Ana Isabel Gonzalez-Gonzalez; gonzalezgonzalez@allgemeinmedizin.uni-frankfurt.de

ABSTRACT

Objective To explore factors that potentially impact external validation performance while developing and validating a prognostic model for hospital admissions (HAs) in complex older general practice patients.

Study design and setting Using individual participant data from four cluster-randomised trials conducted in the Netherlands and Germany, we used logistic regression to develop a prognostic model to predict all-cause HAs within a 6-month follow-up period. A stratified intercept was used to account for heterogeneity in baseline risk between the studies. The model was validated both internally and by using internal-external cross-validation (IECV).

Results Prior HAs, physical components of the health-related quality of life comorbidity index, and medication-related variables were used in the final model. While achieving moderate discriminatory performance, internal bootstrap validation revealed a pronounced risk of overfitting. The results of the IECV, in which calibration was highly variable even after accounting for between-study heterogeneity, agreed with this finding. Heterogeneity was equally reflected in differing baseline risk, predictor effects and absolute risk predictions.

Conclusions Predictor effect heterogeneity and differing baseline risk can explain the limited external performance of HA prediction models. With such drivers known, model adjustments in external validation settings (eg, intercept recalibration, complete updating) can be applied more purposefully.

Trial registration number PROSPERO id: CRD42018088129.

Strengths and limitations of this study

- ► Development of a prognostic model for all-cause hospital admissions using individual participant data yielded clinically plausible predictors.
- ► A significant risk of overfitting in internal validation, and the heterogeneous estimates resulting from internal-external cross-validation as a particular strength, indicated that challenging calibration may have limited external validation performance.
- ► While potential reasons for between-study heterogeneity could be explored, small samples from only four original studies not differentiating between admission causes were obvious limitations.

INTRODUCTION

Growth in the older population raises the frequency of hospital admissions (HAs).[1 2] The increase in HAs reflects not only the ageing population, but also the increased incidence of multiple (chronic) conditions.[3] Moreover, the rising demand for healthcare services also leads to unplanned and potentially preventable HAs, which are an important concern for the healthcare system. These unplanned and potentially preventable HAs can be classified as 'triple fail' events,[4] as they risk being an unpleasant experience for patients, challenging public health and raising health spending.[5] For individual patients, such distressing events make them vulnerable to further adverse events, including falls, increased disabilities and deterioration in health-related quality of life (HRQoL).[6 7] In the context of public health and primary care in particular, physicians have to deal with complex patient needs that entail a higher risk of mismanagement in terms of misdiagnosis and/or mistreatment (ie, medication overuse, misuse or underuse).[8–10] Primary care thus faces the challenge of avoiding such 'triple fail' HA events and instead improving patients' healthcare experiences.[4]

One solution would be to offer timely and appropriate primary care interventions to patients at high risk of HAs. However, in order

to be effective, such preventive interventions should be targeted at those at genuine risk.[11] Numerous prediction models to identify patients at risk of (unplanned) hospitalisations have been developed in various populations.[5 11–16] Several obstacles to good model performance have been identified,[17] but promising methodological advances have neither been able to provide a breakthrough in parametric modelling,[18 19] nor machine learning.[20] External validation in particular has proved to be a major challenge with regard to predictive performance.[21] The model must be able to provide accurate predictions in a new but related situation based on independent data.[22] Generally, model development should balance the number of (meaningful) predictor variables at a reasonably large sample size, while model evaluation also requires enough events when applying the model to a new situation. Even if some of these prerequisites are not fully met, prognostic modelling using individual participant data (IPD) from a meta-analytic (MA) summary of several studies can help to investigate the factors driving external performance.[23] By using IPD-MA, model development can profit from the enlarged casemix variability offered by patients from different healthcare settings, as well as, and more importantly, benefit from the opportunity to simultaneously perform external validation in an approach called internal-external cross-validation (IECV).[24 25] By repeatedly fitting a model to all but one of the IPD trials (ie, training set), IECV mimics the model's application in a new population, while checking predictive performance in the omitted study (ie, test set).

The recently introduced PROPERmed database provides such an IPD framework.[26] Basically, if we want our prediction model to perform well in new, independent patients, between-study heterogeneity with respect to missing values, covariate and endpoint distribution, baseline risks and predictor effects (ie, the associations between predictors and outcome) must be adequately accounted for during model development.[27] While exploring how these key elements drive (external) predictive performance, we are especially concerned with model calibration, the 'Achilles heel' of predictive analytics.[28 29] This is of particular importance because a well-calibrated model is more useful from a clinical perspective than a competing model with better discriminatory performance (by means of the c-statistic or area under the receiver operator characteristics curve, ROC), but worse calibration performance.[30] For example, this can be detrimental in case of systematic overerestimation or underestimation of risks in a new population. Thus, a calibration curve is central to assess calibration: the calibration intercept exposes heterogeneity in baseline risk, and the coefficient of the logistic calibration analysis ('calibration slope') reveals heterogeneous predictor effects.[31] Using an IPD-based model of all-cause HA risk in a way that has previously proved successful,[24] we aim to demonstrate how external validation might be affected by between-study heterogeneity in baseline risk, predictor effects and absolute risk predictions.[27] As an applied clinical example of numerous methods introduced by Steyerberg *et al*,[27] among others, we used IPD methods to predict HA and thus pursued two goals: (1) we expect the findings in our example to help explain the poor external performance of previous prediction models and, looking beyond our particular example, (2) we aim to show that such an approach can guide model developers concerned about poor external performance to choose appropriate methods of model adjustment (eg, intercept recalibration, model updating), if indicated.

METHODS

Source of data and participants

We used harmonised IPD from the PROPERmed database[32] that stem from four trials that qualified for inclusion because they recorded the precise times of study outcomes, namely ISCOPE *(Integrated Systematic Care for Older PEople)*,[33] Opti-Med *(Optimised clinical medication reviews in older people with 'geriatric giants' in general practice)*,[34 35] PRIMUM (*PRIoritising MUltimedication in Multimorbidity in general practices*) [36 37] and RIME (*Reduction of potentially Inappropriate Medication in the Elderly*; Deutsches Register Klinischer Studien-ID, DRKS00003610). Details of the origin and preparation of the source data for the PROPERmed database are described elsewhere.[32] In brief, they were conducted in the Netherlands and Germany between 2009 and 2012 to optimise pharmacological treatment in older chronically ill patients. Three trials (Opti-Med, PRIMUM and RIME) compared a structured medication review consisting of several intervention components with usual care, whereas ISCOPE used a functional geriatric approach to compare usual care with a proactive and integrated plan.

Inclusion criteria for the study participants were identical to our previous work,[38] with patients from general practices being eligible if they were aged 60 years or older, had been diagnosed with at least one chronic condition defined using the O'Halloran list,[39] and had at least one chronic prescription at study baseline (≤2 weeks duration in PRIMUM, ≤2 months in ISCOPE and ≤3 months in Opti-Med and RIME).

Outcome and candidate prognostic variables

As our outcome definition could not distinguish emergency from planned admissions and the source data did not provide information on day and overnight admissions, we defined HAs as a binary outcome for all-cause HAs between baseline and 6-month follow-up. It is worth noting that ISCOPE used a longer follow-up period of 12 months. However, as time-based interactions with predictors did not reveal any statistically significant effect modulation during model development, the resulting potential for confounding can simply be reflected in a different baseline risk.

We had the opportunity to use all PROPERmed variables as candidate predictors, ranging from sociodemographics, lifestyle variables, patient (co)morbidity,

medication, functional status and well-being (eg, HRQoL). The main candidate predictors for this prognostic model were age, sex, living situation, educational level, comorbidities according to the Diederichs list,[40] potentially inappropriate prescriptions according to the European Union (EU) Potentially Inappropriate Medications list,[41] STOPP-START (*STOPP: screening tool of older persons' potentially inappropriate prescriptions; START: screening tool to alert doctors to the right treatment*) criteria,[42] the Dreischulte list,[43] three indices for anticholinergic drug burden,[44–49] harmonised scales indicating depressive symptoms[50–55] or functional decline,[56–58] and two independent subscales from the HRQoL Comorbidity Index.[59–61] In addition to these, we also considered the number of HAs at baseline (ie, during the 12 months before inclusion) as a known strong predictor of future HAs[62] (online supplemental table 1).

Sample size and missing data

Outcome information on HA was complete, while there were sporadically missing values in predictor variables and most importantly, the number of prior HA at baseline was completely missing in the Opti-Med data source. As we expected the number of prior HAs at baseline to be one of the most predictive variable, we chose multilevel multiple imputation[63] to ensure this variable was completely available and, vice versa, to retain all Opti-Med data when this information was systematically missing. We thus considered five iterations of each of six multiple-imputed (MI) datasets,[64] and pooled them according to Rubin's Rules.[65] This procedure was extensively investigated in the PROPERmed database in a previous project[38] with no impact on predictive performance with higher numbers of iterations and imputations. All results were compared with complete-case (CC) analyses, whenever applicable. Missing data and imputation patterns showed reasonable results, whereby this imputation procedure was specifically developed to adjust for within-study and between-study variability (online supplemental figure 1).[66 67] Furthermore, when values were missing systematically, we did not consider the associated candidate prognostic variables in any of original studies (eg, smoking status). Given our final estimate of the c-statistic, sample size, event frequency and number of candidate predictors, we were well aware that this setting would not allow us to obtain an acceptable heuristic shrinkage factor or vice versa, adequate likelihood of a well-performing model.[68]

Methods used in the statistical analysis

Aiming to explore key drivers of external validation performance, we applied a simplified statistical modelling process with a single-imputation dataset (we provided multiple-imputation metrics where applicable), and fitting only one structural model in IECV, and studying heterogeneity using this once defined set of predictor variables.

For model development, we used a fixed-effects logistic regression model with a stratified intercept[27] to conduct IPD analyses and account for between-study heterogeneity[24] in our four eligible studies. The model was thus developed using logistic regression and by adding study indicator variables through the application of effect coding to estimate relative effects with a global average.[69] While these study indicators, along with the basic variables of age and sex, were considered mandatory in model development, all the other 88 prognostic variables were evaluated in a variable selection process that used the so-called Least Absolute Shrinkage and Selection Operator (LASSO)[70] with the 'minCV +1 SE rule'[71] to obtain the sparser models that result from a larger penalty.[72] The final model was derived by using maximum likelihood to refit the model formula,[71] whereby an estimate of overfitting was obtained using internal bootstrap validation.

For model evaluation, we considered the performance metrics of the c-statistic to indicate the discriminatory ability in separating events from non-events by predicted probabilities,[73] calibration intercept to indicated baseline risk specification, calibration slope to indicate predictor effect, calibration-in-the-large (CITL) for a global assessment of the former two,[74] and MA measures for between-study heterogeneity to indicate differences between the four original studies.[75] Internal model validation relied on bootstrap sampling, whereby a model was developed for each of 250 bootstrap samples. The number of samples drawn from each study depended on its sample size thus maintaining the ratio between study participants in bootstrap samples.[76] The c-statistic for the original IPD was derived from these bootstrap models, and arithmetic means were calculated across all bootstrap samples to yield the optimism-corrected c-statistic. To quantify potential optimism, the uniform shrinkage factor was obtained by applying the mean difference in the calibration slopes for each bootstrap model to both the original IPD and in-sample bootstrap performance.[38]

In addition, estimates of generalisability were obtained using IECV, with each study just the once serving as a validation sample for a model developed in the remaining studies.[25] The c-statistic[73] and CITL[74] were the numerical metrics of choice, while calibration plots were visually explored.[30] We thus followed a defined calibration hierarchy[77] that considered CITL to be an important metric for external validation, as well as the calibration slope; the calibration slope was defined as the coefficient of a logistic calibration analysis with cumulated outcomes as the dependent variable and the logit of all predicted risks as the independent variable.[31] Among available options for setting baseline risks (intercept) in validation (test) data,[24] our choice of the average intercept of the IECV training set is considered a conservative option. After extracting c-statistics and CITL estimates at every stage of the IECV loop and obtaining their within-study correlation using a non-parametric bootstrap,[23] the respective estimates were pooled in a random-effects multivariate meta-analysis.[75]

Metrics to explore between-study heterogeneity included the I^2 measure of heterogeneity.[75] In order to

quantify the membership strength of a specific study, we built a multinomial logistic regression model with study indicators as the dependent variables and all selected prognostic variables and the outcome HAs as predictors.[27 74] The c-statistic of this membership model was derived by comparing the predicted probabilities for patients in one specific study with those of patients that were not. Separately, we used pairwise comparisons of the original studies to calculate Pearson correlations between the predictions of study-specific models.[27 74]

All analyses were conducted using the R software environment in V.3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) with the key packages of caret,[78] glmnet (70)(61), metaphor, mice,[64] VIM,[67] pROC[73] and ROCR.[79]

This research study was reported in accordance with the TRIPOD *(Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis)* statement (online supplemental table 2).[80]

Patient and public involvement

Patients or members of the public were not involved in the design, or conduct, or reporting, or dissemination plans of the research.

RESULTS

We included 3804 patients from the available PROPERmed IPD (PRIMUM n=499, Opti-Med n=514, ISCOPE n=1598 and RIME n=1193) (figure 1). Overall, this population had a mean age of 78 years, and 60.3% were female. Based on the chronic conditions defining eligibility and in accordance with the O'Halloran list,[39] 17.9% had been diagnosed with heart failure, 16.4% with chronic obstructive pulmonary disease, 35.7% with non-insulin/dependent diabetes and 12.5% had experienced acute myocardial infarction. In this subset of CC, 598 (21.2 %) patients had been admitted to hospital at least once (table 1).

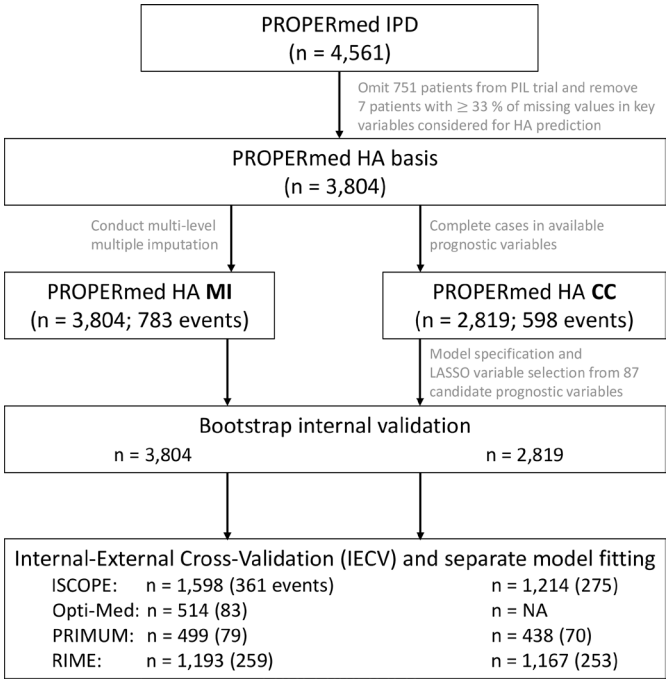

Figure 1 Flow chart and schematic course of action. CC, complete cases; dHRQoL, deterioration of health-related quality of life; HA, hospital admission; IPD, Individual Participant Data; LASSO, Least Absolute Shrinkage and Selection Operator; MI, multiply imputed.

Model development yielded a structural model with seven prognostic variables and study-specific intercepts (table 2). Of the prognostic variables, the number of previous HAs at baseline had the highest effect and partly reflected pronounced casemix variability between the original studies (figure 2A). Similar estimates between CC and MI scenarios supported the use of the imputation procedure to deal with systematically missing numbers of previous HAs at baseline (online supplemental table 3). In internal bootstrap validation, the model achieved an optimism-corrected c-statistic of 0.64 (95% CI 0.62 to 0.67) with a calibration slope of 0.7 (0.6 to 0.83) diverging from one and thus indicating substantial potential for over-fitting. Compared with in-sample metrics for apparent performance, we obtained poor performance, especially in terms of model calibration, when pooling the test study data from each IECV loop (figure 2B,C).

Random-effects meta-analysis of particular studies' test data in the IECV yielded a c-statistic of 0.60 (0.56 to 0.64) and CITL of −0.03 (-0.21 to 0.15). Between-study heterogeneity was striking with I^2 estimates of 50.9% and 61.5 %, respectively. A highly variable performance resulted when the model was applied to each original study separately (figure 3). Among potential drivers of external validation performance, outcome frequencies and thus baseline risks differed strongly, while predicted risks appeared to show a consistent pattern (table 3). Membership c-statistics revealed that the membership model had generally high discriminative ability with respect to identifying the membership of a specific study. This indicates that the predictors and outcome distributions of the studies varied considerably, with patients from the ISCOPE study differing the most. When study-specific models were fitted and applied to the complete IPD, pairwise comparisons revealed moderate to high correlations between the linear predictors of study-specific models (online supplemental figure 2). This suggests that mean estimates involving the entire IPD may enable differences to be balanced out. Similarly, a meta-analysis of single predictor effects from these study-specific models revealed heterogeneity (I^2 measure exceeding 30 %) in age and the number of previous HAs at baseline (online supplemental figure 3).

DISCUSSION

Our applied example takes a pioneering approach to use IPD-based modelling of HAs in general practice in order to expose the challenges of achieving good external validity in such a model. Heterogeneous baseline risks, absolute risk predictions and predictor effects

Table 1 Candidate prognostic variables and statistically significant univariable associations with HAs

Candidate prognostic variable	HAs (complete-case population) No n=2221	HAs (complete-case population) Yes n=598	Descriptive univariable P value
Sociodemographic and lifestyle-related			
Age–mean (SD)	78.2 (6.4)	78.4 (5.8)	0.632
Sex (female)–frequency (%)	1321 (59.5)	330 (55.2)	0.059
Morbidity related			
Cancer–frequency (%)	374 (16.8)	134 (22.4)	0.002
Cerebrovascular disease–frequency (%)	334 (15.0)	113 (18.9)	0.022
Coronary heart disease–frequency (%)	747 (33.6)	239 (40.0)	0.004
Heart failure–frequency (%)	456 (20.5)	169 (28.3)	<0.001
Disease count according to Diederichs*–median (IQR)	3 (3)	4 (3)	<0.001
Medication related			
No of drugs†–median (IQR)	8 (5)	8 (5)	<0.001
Polypharmacy (≥5 drugs)–frequency (%)	1787 (80.5)	503 (84.1)	0.043
Drugs for acid-related disorders–frequency (%)	822 (37.0)	279 (46.7)	<0.001
Drugs for constipation–frequency (%)	161 (7.2)	70 (11.7)	<0.001
Cardiac therapy–frequency (%)	506 (22.8)	171 (28.6)	0.003
Urologicals–frequency (%)	282 (12.7)	107 (17.9)	0.001
Psycholeptics–frequency (%)	272 (12.3)	100 (16.7)	0.004
No of Potentially Inappropriate Medications (PIM) according to the EU-PIM list–Median (IQR)	1 (1)	1 (2)	0.004
Drug Burden Index–median (IQR)	0 (1)	0 (1)	<0.001
Anticholinergic Drug Burden according to Duran–median (IQR)	0 (1)	0 (1)	0.007
Anticholinergic Drug Scale according to Carnahan–median (IQR)	0 (1)	1 (1)	<0.001
STOPP criteria†–median (IQR)	2 (1)	2 (2)	<0.001
STOPP criteria†–frequency (%)	1917 (86.3)	541 (90.5)	0.007
Benzodiazepines–STOPP criteria D5 and K1	191 (8.6)	74 (12.4)	0.005
First generation antihistamines–STOPP criteria D14	29 (1.3)	9 (1.5)	0.708
Hypnotic Z-drugs, for example, zopiclone, zolpidem, zaleplon–STOPP criteria K4	50 (2.3)	23 (3.8)	0.031
Heart failure and prescribed any oral NSAID–Dreischulte B3	64 (2.9)	25 (4.2)	0.109
START criteria‡–median (IQR)	1 (2)	1 (2)	<0.001
START criteria‡–frequency (%)	1325 (59.7)	396 (66.2)	0.004
Documented history of coronary or cerebral vascular disease (aged 85 years and under) and no statin therapy–START criteria A5	230 (10.4)	86 (14.4)	0.006
Heart failure and/or documented coronary artery disease and no ACE inhibitor–START criteria A6	224 (10.1)	81 (13.6)	0.016
Ischaemic heart disease and no beta-blocker–START criteria A7	180 (8.1)	73 (12.2)	0.002
Heart failure and no appropriate beta-blocker (bisoprolol, nebivolol, metoprolol or carvedilol)–START criteria A8	149 (6.7)	64 (10.7)	0.001
Patients taking long-term systemic corticosteroid therapy and no bisphosphonates and vitamin D and calcium–START criteria E2	97 (4.4)	39 (6.5)	0.03
Functional status and well-being related			

Continued

Table 1 Continued

Candidate prognostic variable	HAs (complete-case population) No n=2221	HAs (complete-case population) Yes n=598	Descriptive univariable P value
Functional status–mean (SD)	−0.054 (0.96)	0.093 (0.98)	0.001
Health-related quality of life Comorbidity Index, mental§–median (IQR)	1 (2)	1 (3)	<0.001
Health-related quality of life Comorbidity Index, physical¶–median (IQR)	5 (5)	6 (6)	<0.001
Pain–frequency (%)	1461 (65.8)	427 (71.4)	0.01
Hospital admissions (baseline)**–median (IQR)	0 (0)	0 (1)	<0.001

This table shows candidate prognostic variables stratified according to observed HAs status and univariable associations.
*Twelve conditions were considered over a total of 17 conditions included in the Diederichs list.
†Thirty-two STOPP criteria were considered.
‡Fifteen START criteria were considered.
§Score calculated considering a maximum count of 6 conditions.
¶Score calculated considering a maximum count of 12 conditions.
**ISCOPE, Opti-Med, PRIMUM, RIME.
HAs, hospital admissions; NSAID, non-steroidal anti-inflammatory drugs.

were obvious drivers of the poor external (calibration) performance and should be explored before a particular model is applied to a certain target population. As IPD-based modelling enables this information to be accessed directly, it may be exploited in the modelling process by adapting predictor effects, and ensuring intercepts reflect baseline risks. While pooled average effects may compensate for such differences, separate analysis has revealed how important it is to 'know' as much as possible about the target population to which a model is applied. In the end, a deeper understanding of critical elements can help the developer to choose appropriate methods for model adjustment in the target population, among others intercept re-calibration or (complete) model updating.

Table 2 Final multivariable analysis for HAs after 6 months of follow-up

Prognostic variable	Estimate	SE	P value
Global intercept*	−1.641	0.616	0.008
Age (per year)	−0.010	0.008	0.220
Sex (male)	0.226	0.096	0.016
Medication count†	0.034	0.016	0.032
START criteria count‡	0.080	0.036	0.028
STOPP criteria count§	0.073	0.038	0.056
Physical Component Summary score (PCS) from health-related quality of life Comorbidity Index¶	0.013	0.015	0.373
HAs at baseline**	0.376	0.053	<0.001

*In addition to the study-specific intercept (baseline risks): ISCOPE (0.510), Opti-Med (−0.242), PRIMUM (−0.248), RIME (−0.020).
†Medication count is operationalised as (anatomical therapeutic chemical classification system) 7-digit codes are used for chronic medication as defined per trial including medication for external use.
‡START criteria included START A3, A5-A8, B1, B2, C1, C2, E1-E4, E7 and F1.
§STOPP criteria included STOPP B1-B3, B10, B12, B13, C6, C7, C10, C11, D2, D5-D7, D14, F1, G1, G2, H2-H5, H7, H8, J1-J3, K1-K4 and M1.
¶PCS was calculated according to the modified instrument: maximum count 12 conditions, 47 points.
**Hospital admissions at baseline were absolute number of previous hospital admissions (in the 12 months preceding baseline).
HA, hospital admissions.

IPD modelling with several small data sets for model development and/or model evaluation is promising because larger amounts of data can be used. Regarding our model performance, the small samples from only four studies may not have been large enough, although our performance was similar to previously developed all-cause admission models[19] in its ability to identify well-known prognostic variables (eg, potentially inappropriate prescribing),[81 82] and make corresponding parameter estimates of reasonable magnitude. For example, our model concurs with current research that found prior admissions to be the most relevant prognostic variable, followed by variables related to morbidity and functional disability.[62] In our particular case, morbidity-related measures may also be reflected in the variables used to describe drug utilisation. While well-known diagnoses such as heart failure demonstrated the database's validity by being significantly associated with HAs in univariate analysis (table 1), they did not contribute enough predictive strength to be used in the prognostic model of all-cause HA. This may simply be due to our outcome definition, which did not distinguish between preventable and all-cause HAs. All-cause HAs also included planned visits (which usually exceed 50% of all admissions[83]), which, apart from not having to be predicted, are presumably less dependent on specific factors and thus render such prognostic

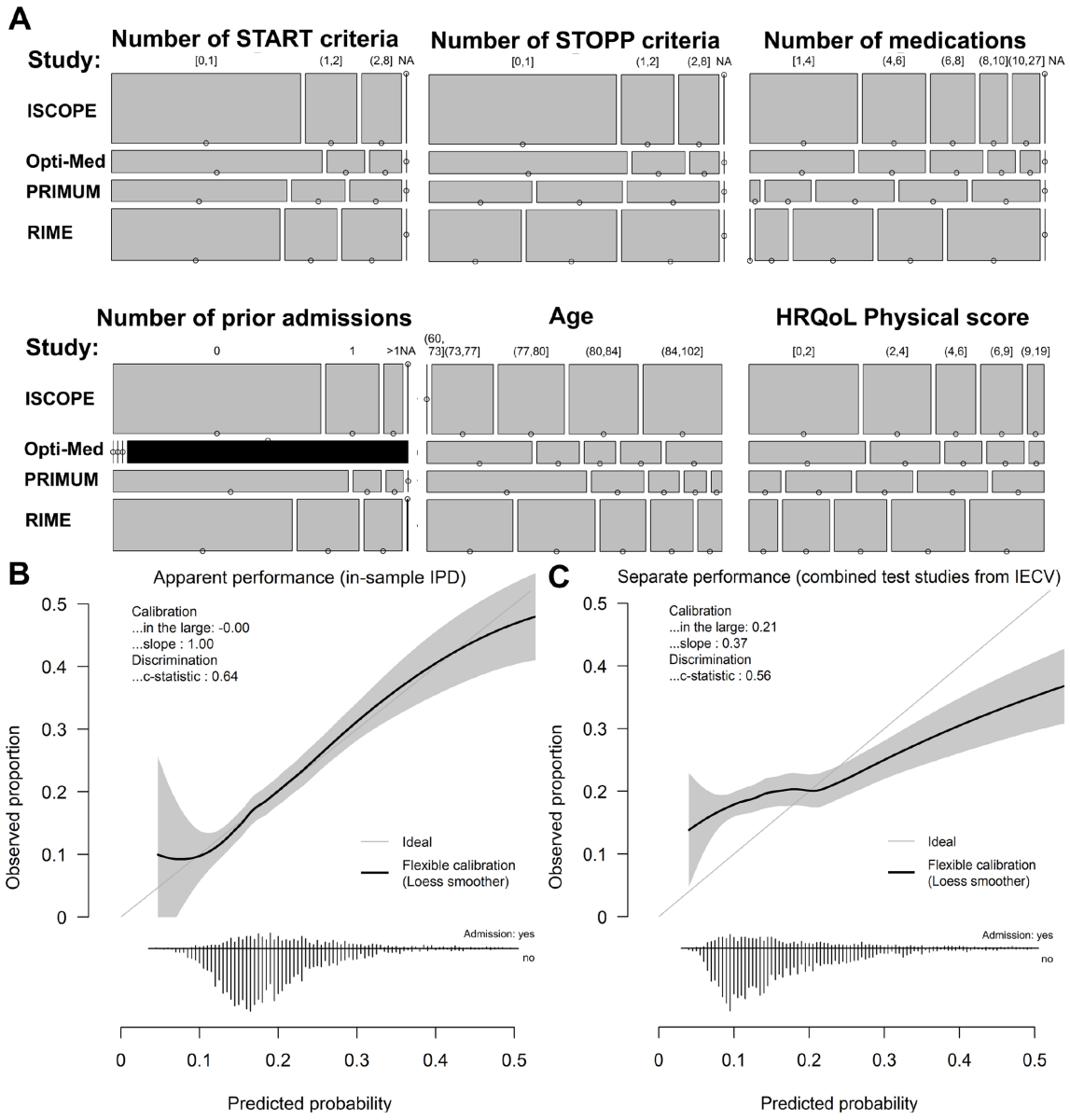

Figure 2 Model development and internal validation. Casemix variability in distributions of prognostic variables is visualised in mosaic plots stratified for the included original studies (area height according to study size; PROPERmed study numbering according to 1: ISCOPE; 2: Opti-med; 4: primum; 5: RIME). The size of the segments represent the number of patients and black areas indicate missing values (A). In calibration plots, predicted probabilities are presented against cumulated observed event proportions for the complete IPD on in-sample application of the HA prediction model (B) and for the combined original study data when used for validation in the IECV (hold-out) (C). HA, hospital admission; IECV, internal-external cross-validation; IPD, individual participant data.

models less sensitive.[81] Above, missing but potentially useful predictor variables that were unavailable for us or predictor misclassifications could also have had a negative impact on our observed performance. Nevertheless, it can be considered as highly favourable that medication-related risk factors are included in our model, as they will facilitate the identification of important issues in interventions targeting medication appropriateness.[8 10] For example, while the number of medications (together with the number of previous HAs) may help in risk stratification, the START and STOPP criteria are conditions that can be directly acted on by changing medication. It thus appears feasible that individual risks can be reduced and the 'Triple Aim' of improving patients' experience of healthcare, advancing public health and lowering per capita costs achieved.[4] As an immediate next step beyond our model, however, we strongly advocate first refining the model's outcome definition to predict preventable HAs.

Using established methods of accounting for between-study heterogeneity,[24] IECV performance was only modest and also expected from the large uniform shrinkage factor of 30% (one minus the optimism-corrected calibration slope). Between-study heterogeneity was moderate to high, and high variation in the results of distinct IECV validation studies clearly emphasised this point. The fact that

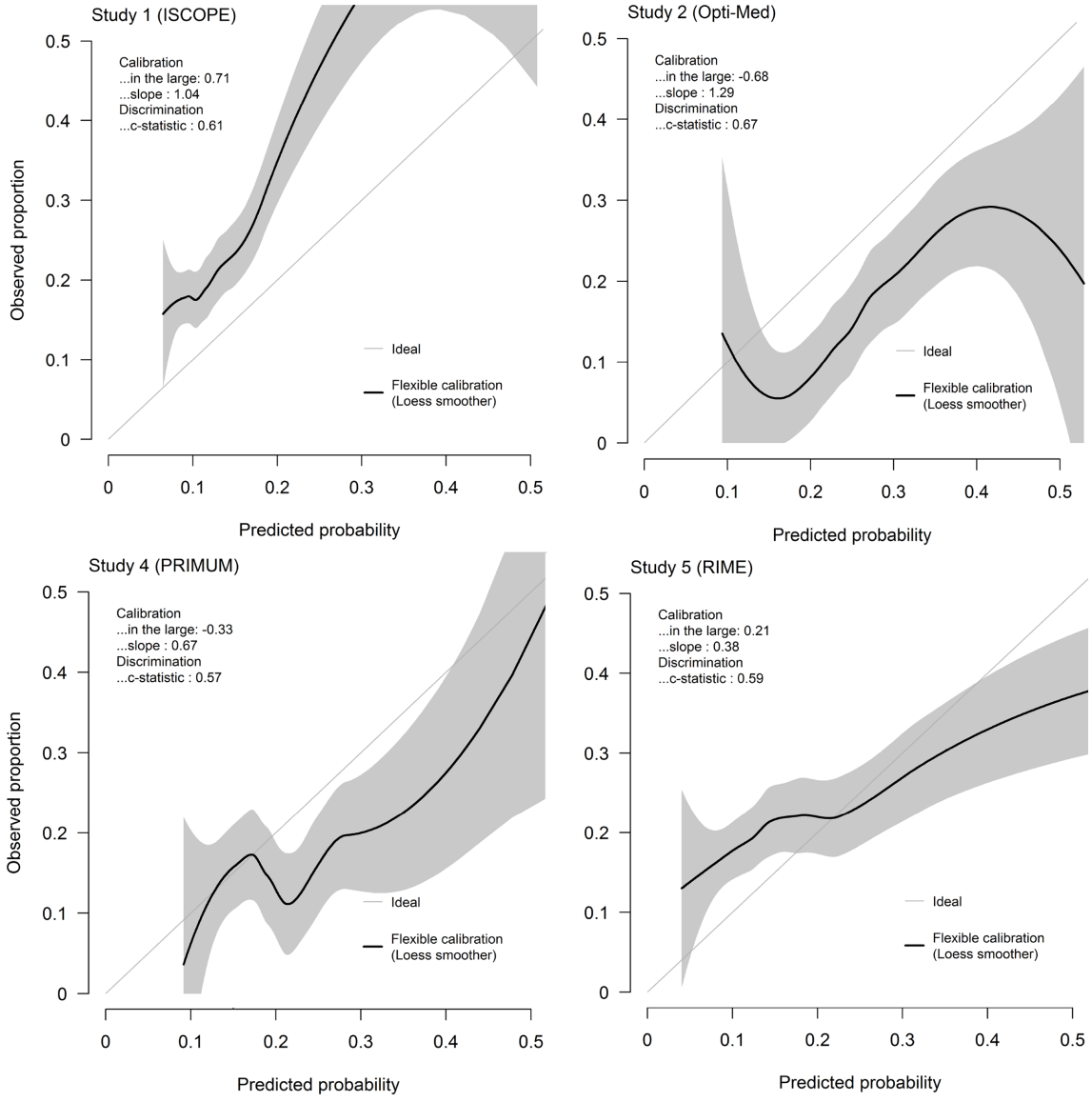

Figure 3 Assessment of between-study heterogeneity. Calibration plots are obtained from each data subset when a particular original study served as the validation sample in the IECV. IECV, internal-external cross-validation.

the global intercept also indicated pronounced heterogeneity in the original studies suggests that the current set of predictors did not explain variability to the extent necessary for the design of a better performing prediction model (online supplemental figure 3). The study indicators alone clearly did not adequately reflect the baseline risks of populations from different healthcare systems, which may also mean that the 'right' prognostic variables for predicting all-cause HAs were not available, or not to the necessary degree informative.

Table 3 Between-study heterogeneity

		Baseline risk	Linear predictor (=predicted absolute risks)		
Study no	Study name	Admission proportion	Mean	SD	Membership C
1	ISCOPE	0.23	−1.27	−0.46	0.84
2	Opti-Med	0.16	−1.71	−0.28	0.69
4	PRIMUM	0.16	−1.72	−0.52	0.80
5	RIME	0.22	−1.35	−0.33	0.80

Heterogeneity between original studies is described in terms of baseline risk (proportion of participants with hospital admissions), casemix distribution with respect to predicted risks, and the discriminative ability of the membership model to identify membership of a specific study.

Further limitations first relate to the sample sizes needed in model development[68] and validation,[84] as a larger sample size would certainly have been desirable. For instance, in the IECV loop, for which validation data came from original individual studies, we could not meet the requirement of the suggested 100 events for a reliable assessment of predictive performance,[85 86] or the required minimum of 200 patients with and 200 patients without a condition, which would be needed to generate precise calibration curves.[77] The ability to predict unplanned and preventable HAs would have strengthened the potential clinical usefulness of the model. Nevertheless, currently available IPD from PROPERmed do not prevent us from drawing conclusions for future research, which was our primary goal and also the reason for several simplifications to enhance interpretability.

CONCLUSION

Based on PROPERmed IPD-MA, we have illustrated how predictor effect heterogeneity and varying baseline risks can limit the external performance of HA prediction models. Likewise, this approach proved that IPD-based modelling can project external performance and thus help developers addressing the potentially challenging performance after exploring its key drivers. If indicated by IPD, a model might be more purposefully improved when transferred to a new setting by adjusting baseline risks (ie, intercept recalibration) or additionally its predictor effects (ie, model updating).

Author affiliations
[1]Department of Clinical Pharmacology & Pharmacoepidemiology, Heidelberg University, Heidelberg, Baden-Württemberg, Germany
[2]Institute of General Practice, Goethe University, Frankfurt am Main, Hessen, Germany
[3]Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC), Madrid, Spain
[4]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands
[5]School of CAPHRI, Department of Family Medicine, Maastricht University, Maastricht, The Netherlands
[6]Department of General Practice and Elderly Care Medicine, Amsterdam UMC, Vrije Universiteit, Amstedarm Public Health Research Institute, Amsterdam, The Netherlands
[7]Chair of Geriatrics and Gerontology, University Clinic Eppendorf, Hamburg, Germany
[8]Institute for Evidence in Medicine (for Cochrane Germany Foundation), Medical Center-University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[9]Department of Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Bochum, Nordrhein-Westfalen, Germany
[10]Techniker Krankenkasse (TK), Hamburg, Germany
[11]Centre for Prognosis Research, School of Primary Care Research, Community and Social Care, Keele University, Keele, UK
[12]Nuffield Department of Primary Care, University of Oxford, Oxford, UK
[13]Centre for Research in Evidence-Based Practice, Bond University, Robina, Queensland, Australia
[14]Department of General Practice and Family Medicine, Medical Faculty OWL, University of Bielefeld, Bielefeld, Germany

Twitter Joerg J Meerpohl @meerpohl

Acknowledgements The authors would like to thank all participating local data managers (Sandra Rauck, Mascha Twellaar, Karin Aretz, Antonio Fenoy, and Kiran Chapidi). We would also like to thank Phillip Elliott for editing the manuscript.

Contributors JB, MvdA, UT, WEH, HJT, DB-L, PE, GK, JJM, DKdG, RP, PG, FMG, ADM and CM contributed to the design of the PROPERmed study. CM is the guarantor. ADM and AIG-G wrote the first draft of the manuscript. AIG-G and TSD developed the harmonised PROPERmed database; KMAS, HR and BF provided support. ADM performed the statistical analysis; RP, KIES and HR provided support. All authors contributed to the manuscript and agreed on its publication. All authors are members of the PROPERmed project being involved from the very beginning with significant contributions to conceptualisation, data harmonisation, design of analysis and interpretation of results. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding This work was supported by the German Innovation Fund in accordance with § 92a (2) Volume V of the Social Insurance Code (§ 92a Abs. 2, SGB V - Fünftes Buch Sozialgesetzbuch), grant number: 01VSF16018. ADM is sponsored by the Physician-Scientist Programme of Heidelberg University, Faculty of Medicine. Rafael Perera receives funding from the NIHR Oxford Biomedical Research Council (BRC), the NIHR Oxford Medtech and In-Vitro Diagnostics Co-operative (MIC), the NIHR Applied Research Collaboration (ARC) Oxford and Thames Valley, and the Oxford Martin School. KIES is sponsored by the National Institute for Health Research School for Primary Care Research (NIHR SPCR Launching Fellowship).

Disclaimer The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The ethics commission of the medical faculty of the Johann Wolfgang Goethe University, Frankfurt / Main confirmed that no extra vote was necessary for the anonymous use of data from the PROPERmed IPD-MA (13/07/2017). All included studies were separately approved by the relevant ethics commissions as follows: ISCOPE: The Medical Ethical Committee of Leiden University Medical Center approved the study (date: 30.06.2009, reference: P09.096). Opti-Med: The Medical Ethics Committee of the VU University Medical Centre Amsterdam approved the study (date: 12.01.2012, reference: 2011/408). PIL: The Medical Ethics Review Board Atrium-Orbis-Zuyd approved the study (date: 15.12.2009, reference: 09-T-72 NL3037.096.09). PRIMUM: The Ethics Commission of the Medical Faculty of the Johann Wolfgang Goethe University, Frankfurt / Main approved the study (date: 20/05/2010, reference: E 46/10). RIME: The Ethics Commission of the University Witten / Herdecke approved the study (date: 28.02.2012, reference: 147/2011).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information. Source data originate from separate primary studies and can potentially be requested for anonymous use from the PROPERmed IPD-MA database.

ORCID iDs
Ana Isabel Gonzalez-Gonzalez http://orcid.org/0000-0002-1707-0596
Marjan van den Akker http://orcid.org/0000-0002-1022-8637
Christiane Muth http://orcid.org/0000-0001-8987-182X

REFERENCES

1 Schuur JD, Venkatesh AK. The growing role of emergency departments in hospital admissions. *N Engl J Med* 2012;367:391–3.
2 Wittenberg R, Sharpin L, McCormick B, *et al*. The ageing Society and emergency hospital admissions. *Health Policy* 2017;121:923–8.
3 Barnett K, Mercer SW, Norbury M, *et al*. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;380:37–43.
4 Lewis G, Kirkham H, Duncan I, *et al*. How health systems could avert 'triple fail' events that are harmful, are costly, and result in poor patient satisfaction. *Health Aff* 2013;32:669–76.
5 Wallace E, Stuart E, Vaughan N, *et al*. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Med Care* 2014;52:751–65.
6 Covinsky KE, Palmer RM, Fortinsky RH, *et al*. Loss of independence in activities of daily living in older adults hospitalized with medical illnesses: increased vulnerability with age. *J Am Geriatr Soc* 2003;51:451–8.
7 Keeble E, Roberts HC, Williams CD, *et al*. Outcomes of hospital admissions among frail older people: a 2-year cohort study. *Br J Gen Pract* 2019;69:e555–60.
8 Haefeli WE, Meid AD. Pill-count and the arithmetic of risk: evidence that polypharmacy is a health status marker rather than a predictive surrogate for the risk of adverse drug events. *Int J Clin Pharmacol Ther* 2018;56:572–6.
9 L Reed R, Isherwood L, Ben-Tovim D. Why do older people with multi-morbidity experience unplanned hospital admissions from the community: a root cause analysis. *BMC Health Serv Res* 2015;15:525.
10 Meid AD, Lampert A, Burnett A, *et al*. The impact of pharmaceutical care interventions for medication underuse in older people: a systematic review and meta-analysis. *Br J Clin Pharmacol* 2015;80:768–76.
11 Alonso-Morán E, Nuño-Solinis R, Onder G, *et al*. Multimorbidity in risk stratification tools to predict negative outcomes in adult population. *Eur J Intern Med* 2015;26:182–9.
12 Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011;306:1688.
13 Marcusson J, Nord M, Dong H-J, *et al*. Clinically useful prediction of hospital admissions in an older population. *BMC Geriatr* 2020;20:95.
14 Coleman EA, Wagner EH, Grothaus LC, *et al*. Predicting hospitalization and functional decline in older health plan enrollees: are administrative data as accurate as self-report? *J Am Geriatr Soc* 1998;46:419–25.
15 Haas LR, Takahashi PY, Shah ND, *et al*. Risk-Stratification methods for identifying patients for care coordination. *Am J Manag Care* 2013;19:725–32.
16 Crane SJ, Tung EE, Hanson GJ, *et al*. Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits: the elders risk assessment index. *BMC Health Serv Res* 2010;10:338.
17 Wallace E, Johansen ME. Clinical prediction rules: challenges, barriers, and promise. *Ann Fam Med* 2018;16:390–2.
18 Meid AD, Groll A, Schieborr U, *et al*. How can we define and analyse drug exposure more precisely to improve the prediction of hospitalizations in longitudinal (claims) data? *Eur J Clin Pharmacol* 2017;73:373–80.
19 Meid AD, Groll A, Heider D, *et al*. Prediction of drug-related risks using clinical context information in longitudinal claims data. *Value Health* 2018;21:1390–8.
20 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
21 Wallace E, McDowell R, Bennett K, *et al*. External validation of the probability of repeated admission (PRA) risk prediction tool in older community-dwelling people attending general practice: a prospective cohort study. *BMJ Open* 2016;6:e012336.
22 Altman DG, Vergouwe Y, Royston P, *et al*. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
23 Snell KIE, Hua H, Debray TPA, *et al*. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2016;69:40–50.
24 Debray TPA, Moons KGM, Ahmed I, *et al*. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32:3158–80.
25 Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med* 2004;23:907–26.
26 González-González AI, Dinh TS, Meid AD, *et al*. Predicting negative health outcomes in older general practice patients with chronic illness: rationale and development of the PROPERmed harmonized individual participant data database. *Mech Ageing Dev* 2021;194:111436.
27 Steyerberg EW, Nieboer D, Debray TPA, *et al*. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: an overview and illustration. *Stat Med* 2019;38:4290–309.
28 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
29 Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics. *JAMA* 2018;320:27.
30 Van Calster B, Vickers AJ. Calibration of risk prediction models. *Med Decis Making* 2015;35:162–9.
31 Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *J Clin Epidemiol* 2020;118:93–9.
32 González-González AI, Dinh TS, Meid AD, *et al*. Predicting negative health outcomes in older general practice patients with chronic illness: rationale and development of the PROPERmed harmonized individual participant data database. *Mech Ageing Dev* 2021;194:111436.
33 Blom J, den Elzen W, van Houwelingen AH, *et al*. Effectiveness and cost-effectiveness of a proactive, goal-oriented, integrated care model in general practice for older people. A cluster randomised controlled trial: Integrated Systematic Care for older People--the ISCOPE study. *Age Ageing* 2016;45:30–41.
34 Willeboordse F, Schellevis FG, Chau SH, *et al*. The effectiveness of optimised clinical medication reviews for geriatric patients: Opti-Med a cluster randomised controlled trial. *Fam Pract* 2017;34:437–45.
35 Willeboordse F, Hugtenburg JG, van Dijk L, *et al*. Opti-Med: the effectiveness of optimised clinical medication reviews in older people with 'geriatric giants' in general practice; study protocol of a cluster randomised controlled trial. *BMC Geriatr* 2014;14:116.
36 Muth C, Harder S, Uhlmann L, *et al*. Pilot study to test the feasibility of a trial design and complex intervention on prioritising MUltimedication in multimorbidity in general practices (PRIMUMpilot). *BMJ Open* 2016;6:e011613.
37 Muth C, Uhlmann L, Haefeli WE, *et al*. Effectiveness of a complex intervention on prioritising Multimedication in multimorbidity (primum) in primary care: results of a pragmatic cluster randomised controlled trial. *BMJ Open* 2018;8:e017740.
38 González-González AI, Meid AD, Dinh TS, *et al*. A prognostic model predicted deterioration in health-related quality of life in older patients with multimorbidity and polypharmacy. *J Clin Epidemiol* 2021;130:1–12.
39 O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. *Fam Pract* 2004;21:381–6.
40 Diederichs C, Berger K, Bartels DB. The measurement of multiple chronic diseases--a systematic review on existing multimorbidity indices. *J Gerontol A Biol Sci Med Sci* 2011;66:301–11.
41 Renom-Guiteras A, Meyer G, Thürmann PA. The EU(7)-PIM list: a list of potentially inappropriate medications for older people consented by experts from seven European countries. *Eur J Clin Pharmacol* 2015;71:861–75.
42 O'Mahony D, O'Sullivan D, Byrne S, *et al*. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. *Age Ageing* 2015;44:213–8.
43 Dreischulte T, Donnan P, Grant A, *et al*. Safer prescribing--a trial of education, informatics, and financial incentives. *N Engl J Med* 2016;374:1053–64.
44 Carnahan RM, Lund BC, Perry PJ, *et al*. The anticholinergic drug scale as a measure of drug-related anticholinergic burden: associations with serum anticholinergic activity. *J Clin Pharmacol* 2006;46:1481–6.
45 Carnahan RM, Lund BC, Perry PJ, *et al*. The relationship of an anticholinergic rating scale with serum anticholinergic activity in elderly nursing home residents. *Psychopharmacol Bull* 2002;36:14–19.
46 Hilmer SN, Mager DE, Simonsick EM, *et al*. A drug burden index to define the functional burden of medications in older people. *Arch Intern Med* 2007;167:781.
47 Cao Y-J, Mager DE, Simonsick EM, *et al*. Physical and cognitive performance and burden of anticholinergics, sedatives, and ACE inhibitors in older women. *Clin Pharmacol Ther* 2008;83:422–9.
48 Hilmer SN, Mager DE, Simonsick EM, *et al*. Drug burden index score and functional decline in older people. *Am J Med* 2009;122:e1-2:1142–9.

49 Durán CE, Azermai M, Vander Stichele RH. Systematic review of anticholinergic risk scales in older adults. *Eur J Clin Pharmacol* 2013;69:1485–96.
50 Sheikh JI, Yesavage JA, Brooks JO, *et al*. Proposed factor structure of the geriatric depression scale. *Int Psychogeriatr* 1991;3:23–8.
51 Yesavage JA, Brink TL, Rose TL, *et al*. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res* 1982;17:37–49.
52 Hoyl MT, Alessi CA, Harker JO, *et al*. Development and testing of a five-item version of the geriatric depression scale. *J Am Geriatr Soc* 1999;47:873–8.
53 Aaronson NK, Muller M, Cohen PD, *et al*. Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. *J Clin Epidemiol* 1998;51:1055–68.
54 Gandek B, Ware JE, Aaronson NK, *et al*. Cross-Validation of item selection and scoring for the SF-12 health survey in nine countries: results from the IQOLA project. International quality of life assessment. *J Clin Epidemiol* 1998;51:1171–8.
55 Ware J, Kosinski M, Keller SD. A 12-Item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34:220–33.
56 Palmer M, Harley D. Models and measurement in disability: an international review. *Health Policy Plan* 2012;27:357–64.
57 Saliba D, Elliott M, Rubenstein LZ, *et al*. The vulnerable elders survey: a tool for identifying vulnerable older people in the community. *J Am Geriatr Soc* 2001;49:1691–9.
58 Isaacs B. *An introduction to geriatrics*. London: Bailliere, Tindall & Cassell, 1965.
59 Mukherjee B, Ou H-T, Wang F, *et al*. A new comorbidity index: the health-related quality of life comorbidity index. *J Clin Epidemiol* 2011;64:309–19.
60 Ou H-T, Mukherjee B, Erickson SR, *et al*. Comparative performance of comorbidity indices in predicting health care-related behaviors and outcomes among Medicaid enrollees with type 2 diabetes. *Popul Health Manag* 2012;15:220–9.
61 Cheng L, Cumber S, Dumas C, *et al*. Health related quality of life in pregeriatric patients with chronic diseases at urban, public supported clinics. *Health Qual Life Outcomes* 2003;1:63.
62 García-Pérez L, Linertová R, Lorenzo-Riera A, *et al*. Risk factors for hospital readmissions in elderly patients: a systematic review. *QJM* 2011;104:639–51.
63 Jolani S, Debray TPA, Koffijberg H, *et al*. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice. *Stat Med* 2015;34:1841–63.
64 Buuren Svan, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in *R*. *J Stat Softw* 2011;45.
65 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Ltd, 1987.
66 Zhang Z. Missing data exploration: highlighting graphical presentation of missing pattern. *Ann Transl Med* 2015;3:356.
67 Kowarik A, Templ M. Imputation with the R package VIM. *J Stat Softw* 2016;74.
68 Riley RD, Snell KI, Ensor J, *et al*. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276–96.
69 Te Grotenhuis M, Pelzer B, Eisinga R, *et al*. When size matters: advantages of weighted effect coding in observational studies. *Int J Public Health* 2017;62:163–7.
70 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
71 Thao LTP, Geskus R. A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biom J* 2019;61:343–56.
72 Lipkovich IA, Dmitrienko A, Ralph B. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017;36:136–96.
73 Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
74 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
75 Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36.
76 Efron B, Tibshirani R. *An introduction to the bootstrap*. CRC Boca Raton London New York Washington, D.C.: Chapman & Hall, 1993.
77 Van Calster B, Nieboer D, Vergouwe Y, *et al*. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167–76.
78 Kuhn M. Building predictive models in *R* using the caret package. *J Stat Softw* 2008;28.
79 Sing T, Sander O, Beerenwinkel N, *et al*. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;21:3940–1.
80 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.
81 van der Stelt CAK, Vermeulen Windsant-van den Tweel AMA, Egberts ACG, *et al*. The association between potentially inappropriate prescribing and medication-related hospital admissions in older patients: a nested case control study. *Drug Saf* 2016;39:79–87.
82 Pérez T, Moriarty F, Wallace E, *et al*. Prevalence of potentially inappropriate prescribing in older people in primary care and its association with hospital admission: longitudinal study. *BMJ* 2018;363:k4524.
83 Schöpke T, Plappert T. Kennzahlen von notaufnahmen in deutschland. *Notfall + Rettungsmedizin* 2011;14:371–8.
84 Steyerberg EW. Validation in prediction research: the waste by data splitting. *J Clin Epidemiol* 2018;103:131–3.
85 Vergouwe Y, Steyerberg EW, Eijkemans MJC, *et al*. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475–83.
86 Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016;76:175–82.