Journal of Clinical Epidemiology 172 (2024) 111387

Journal of Clinical Epidemiology

COMMENTARY

Don't be misled: 3 misconceptions about external validation of clinical prediction models

Hannah M. la Roi-Teeuw[a,*], Florien S. van Royen[a,1], Anne de Hond[b,1], Anum Zahra[b,1], Sjoerd de Vries[c,d,1], Richard Bartels[c,e], Alex J. Carriero[b], Sander van Doorn[a], Zoë S. Dunias[b], Ilse Kant[c], Tuur Leeuwenberg[b], Ruben Peters[c], Laura Veerhoek[c], Maarten van Smeden[b,e], Kim Luijken[b]

[a]*Department of General Practice and Nursing Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands*

[b]*Department of Epidemiology and Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands*

[c]*Department of Digital Health, University Medical Center Utrecht, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands*

[d]*Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, The Netherlands*

[e]*Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584CX, Utrecht, The Netherlands*

Accepted 2 May 2024; Published online 8 May 2024

Abstract

Clinical prediction models provide risks of health outcomes that can inform patients and support medical decisions. However, most models never make it to actual implementation in practice. A commonly heard reason for this lack of implementation is that prediction models are often not externally validated. While we generally encourage external validation, we argue that an external validation is often neither sufficient nor required as an essential step before implementation. As such, any available external validation should not be perceived as a license for model implementation. We clarify this argument by discussing 3 common misconceptions about external validation. We argue that there is not one type of recommended validation design, not always a necessity for external validation, and sometimes a need for multiple external validations. The insights from this paper can help readers to consider, design, interpret, and appreciate external validation studies.

Keywords: External validation; Prediction model; Clinical algorithm; Machine learning; Regression modelling; Study design; Internal validation; Model updating; Clinical prediction model; Artificial intelligence

1. Introduction

A clinical prediction model (CPM) is a model that aims to predict prognostic or diagnostic outcomes in a particular group of individuals, often to support medical decision-making. CPMs exist in varying modalities, ranging from simple ''rules of thumb'' to advanced risk prediction models that are based on artificial intelligence (AI) [1,2]. Two examples of diagnostic CPMs used throughout this article are the Wells score [3] and UrinCheck [4]. The Wells score is a relatively simple score to assess the risk of pulmonary embolism that is widely used in various clinical settings (Box 1). UrinCheck is an AI-based model that predicts urinary tract infection in patients with a suspicion of such an infection and is intended for local use at the hospital at which it was developed (Box 2).

Between the development of a CPM and its implementation in clinical practice, it is necessary to evaluate whether the CPM performance is deemed adequate (Fig) [5]. Currently, most developed CPMs never make it to clinical implementation for several reasons [6]. One potential reason is that well-validated CPMs are scarce [7–9]. To

Funding: No funding was received for this project.

[1] Authors contributed equally.

* Corresponding author. Department of General Practice and Nursing Science, Julius Center UMC Utrecht, Stratenum 6.131, 3508 GA Utrecht, The Netherlands.

E-mail address: h.m.teeuw@umcutrecht.nl (H.M. la Roi-Teeuw).

https://doi.org/10.1016/j.jclinepi.2024.111387

What is new?

- External validation is often perceived as a crucial step to bring clinical prediction models from development to implementation.
- We discuss 3 common misconceptions about external validation, using 2 clinical examples, to illustrate that an external validation can be neither sufficient nor required depending on the intended use of the clinical prediction model.
- The insights from this article can help readers to consider, design, interpret, and appreciate external validation studies.

resolve this, some journals endorse external validation as a requirement for publication of CPM development studies. However, as we will set out in this article, in some situations, the intended CPM use does not require external validation or the intended use and external validation are not aligned. Although external validation likely increases confidence in the CPM, it actually provides little information about CPM performance upon implementation in such cases [10].

In addition, the terms ''internal validation'' and ''external validation'' raise confusion. Internal validation is commonly defined as an evaluation of the CPM's performance in a sample from the same population that was included in model development (optimism-corrected or cross-validated performance measures should be used rather than apparent model performance to prevent overoptimism) [11,12]. External validation commonly refers to CPM performance evaluation in data sampled from a different population [11,12]. In practice, however, it is not always clear whether data were sampled from the same population or not—that may merely rely on semantics (eg, consider a validation in a ''test'' dataset from a train-test split of the data)—and therefore these terms are sometimes confused [12,13]. Focus on the semantics of definitions may distract from the actual purpose that different types of validation serve. Different CPMs will require different types of validation for different purposes. Therefore, general statements about external validation may lead to misconceptions if they are interpreted in a ''one-size-fits-all'' manner. We think that a more aim-focused perspective on validation may facilitate discussions about the necessity, proper design, and interpretation of validation studies with different stakeholders.

In this article, we highlight 3 common misconceptions about external validations. We use examples from the Wells score and UrinCheck to illustrate how the results of external validations can be informative in relevant scenarios, but also misleading when wrongly applied or interpreted. Building on these ideas, we close with 3 considerations that can help the reader in designing and appreciating validation studies (Fig).

Box 1 Wells score

The diagnostic Wells score indicates the risk of current pulmonary embolism [3]. It is widely used to exclude a pulmonary embolism in patients suspected of this diagnosis in primary care. Clinicians calculate a score from 7 simple clinical predictors and order a D-dimer blood test if the score is 4 or lower. The patient needs referral for further diagnostic work-up if the score is 4.5 or higher, or if the D-dimer result is above a threshold. The performance of the Wells score in safely excluding pulmonary embolism has been externally validated in different settings. For example, an individual patient data meta-analysis showed that the failure rate (ie, the proportion of patients with missed pulmonary embolism among those classified as ''pulmonary embolism excluded'' by the Wells score) among nursing home residents or in hospitalized patients was 1.8% (95% CI: 0.7−4.9). This was well above the average failure rate observed in primary health care, which was 0.13% (95% CI 0.03−0.62) [14].

Box 2 UrinCheck

UrinCheck is a local diagnostic clinical decision support system intended to predict current urinary tract infection in patients suspected of having such an infection, aimed at reducing unnecessary early-initiated antibiotic prescriptions [4]. In patients with a leukocyte esterase−positive or nitrite-positive urinalysis, the presence of a urinary tract infection is predicted by a semisupervised machine learning algorithm trained on expert-labeled and unlabeled data from electronic health records of the local hospital in which UrinCheck was developed and intended to be used. Upon internal validation, the model was estimated to predict infection with 76.8% accuracy (95% CI: 75.8−77.8), and it was estimated that implementing the system into routine care could potentially reduce the number of inappropriate antibiotic prescriptions by up to 15.2%.

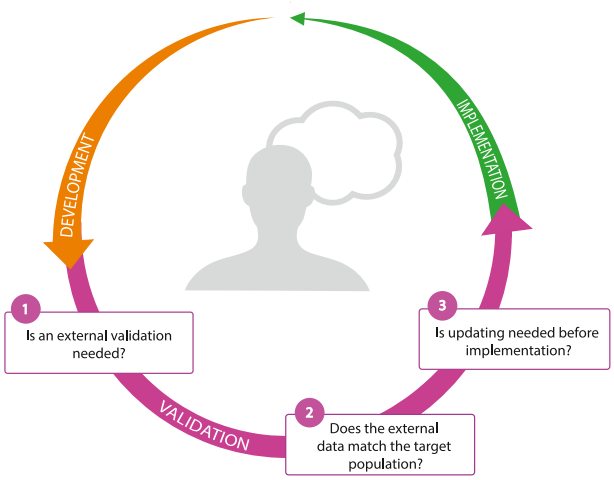

Figure. Three considerations about external validation during different stages of the validation phase in the clinical prediction model lifecycle.

2. Misconceptions

2.1. Misconception 1: "External validation provides stronger evidence of model performance than internal validation"

The first misconception is based on the notion that external validation is a broader, more encompassing proof of validity than internal validation. This belief is maintained by journals and funding agencies who request external validations (often on data from a different location) to corroborate the validity of the CPM's performance, suggesting that an external validation yields a better indication of a CPM's performance than the internal validation of the same model [15]. However, such a hierarchical comparison between internal and external validation does not acknowledge the different objectives that both validations serve. In fact, either or both validation types may be required to assess a CPM's performance in its intended setting of implementation [12].

Internal validation can provide—besides the likely optimistic apparent performance—an optimism-corrected performance estimate by using different data samples from the same underlying population (eg, using cross-validation or bootstrapping) [11,15]. The 95% CIs for these estimates also inform about the uncertainty around the true performance of the CPM in this population. Internal validation can thus directly be informative if the CPM is intended to be used in the same population as used for model development. External validation, on the other hand, measures the transportability of a CPM to a setting different from model development [10]. External validation may, for example, assess a CPM's performance in another hospital, in another target population, under different methods of data collection or at a later point in time when some aspects of the population may have changed [11,15].

The assessment of model performance should be aligned with its intended use [10,16]. Whereas internal validation is generally recommended to identify poor reproducibility of model development in the development population (perhaps due to overfitting to a specific data sample from that population), the need for external validation depends on the type of transportability that is potentially required [12]. Notably, not all models require transportability. For example, for the UrinCheck system, clinicians may want to use the CPM to diagnose patients only in the hospital in which this CPM was developed. In this case, it may suffice to internally validate the CPM performance [2]. In fact, suboptimal external validation results from validation in another hospital (eg, because this hospital uses different equipment) may erode trust in the CPM with clinicians at the original hospital, which would be unjustified.

In short, external validation is not necessarily superior to internal validation. These 2 broad categories of validation are designed for different performance testing goals (ie, internal validity vs transportability). Accordingly, validations should be designed in line with the CPM's intended purpose and not to just "check the box" of having an external validation.

2.2. Misconception 2: "External validation proves the generalizability of model performance"

Another misconception arises when "robust" performance upon external validation is seen as a general indicator of a CPM's validity. Articles may contain statements like "the CPM is a validated tool" or "the CPM has high predictive value." Such statements suggest a sort of "overall generalizability" because the CPM retained good performance despite substantial heterogeneity between the development and (multiple) validation populations. We do recognize that multiple external validations can provide information about CPM performance stability across settings. Such "performance heterogeneity" assessments may particularly be informative for CPMs that are widely used in multiple settings [10,13]. However, for local clinical implementation, the CPM's performance in multiple populations or even *any* population does not necessarily provide insight into its performance in the *target* population for its intended use [10]. If external validation is required, the validation data should ideally be representative of the target population and the intended use of the CPM.

The Wells score is a prime example of a CPM that is externally validated and used in different populations. However, the mere fact that the score "is externally validated" should not be considered a license to use the score in these populations. External validation studies in fact illustrated that the Wells score performed well in the primary care setting, but that the percentage of missed pulmonary embolism diagnoses was unacceptable in hospitalized and nursing home care (Box 1) [14]. This study exemplifies that the results of external validations are specific to the

target population and setting of use, and that generalization to other settings warrants caution.

A well-known problem in current practice is that many external validations are ''off-target'' because the choice of data for these studies is driven by convenience (eg, using readily available datasets) rather than the intended CPM use [10]. For example, an important intended use of the Wells score is to assist general practitioners in their decision whether to refer patients suspected of pulmonary embolism to the hospital for diagnostic imaging or not. One could perform an external validation of the Wells score on routine care hospital data of all patients who were referred for diagnostic imaging for suspected pulmonary embolism. Such data could be readily available from hospital records. However, these data are clearly not representative of the target population for intended use, since the referred patients represent only a very selective subset of the target population. Hence, such an external validation study could provide misleading results to general practitioners [17].

Of course, it is not always feasible to directly obtain validation results for every specific intended setting. For instance, a geriatrician who wants to use the Wells score in a nursing home might find it challenging to identify external validation studies that are relevant to the intended target population. Older people in nursing homes constitute a heterogeneous population [18]. Validation studies may also be sparse because of several challenges in obtaining data from nursing home settings [18]. In general, conducting a local validation at each site where a CPM is to be used can be quite impractical. Therefore, broadly used CPMs like the Wells score may require multiple external validation studies capturing different aspects of the intended target population and settings, to demonstrate their transportability. In the example above, it might for instance be informative to have validations in an older population outside the nursing home (eg, in an older primary care population) or in older patients with specific comorbidities. Note that these represent not just multiple external validations in ''different'' settings to imply ''general robustness'' of CPM performance, but still specific validation studies targeted at different aspects for which transportability is sought.

In summary, when implementation of a CPM in a particular clinical setting is desirable, external validation studies need to reflect the target population and total context of its intended use. The use of datasets that do not reflect the clinically relevant target setting could provide misleading results [10].

2.3. Misconception 3: "External validation is a single-step process before implementation in clinical practice"

A third common misconception is the belief that once a model has been externally validated in a population that aligns with its intended use, and has demonstrated good performance, the model is ''good to go'' for clinical implementation. However, in some circumstances, it is desirable to update the CPM in the local setting for optimal performance. Also, validations at one point in time—even if the studies are of high quality—are often insufficient. Characteristics of the population or care routines (and therefore CPM performance) may change over time. Such changes may even be the desired result of a CPM's use (eg, fewer urinary tract infections with resistant bacteria due to fewer antibiotic prescriptions after implementation of UrinCheck) [19]. Some authors therefore even argue that there is no such a thing as a validated model [20].

First, it is important to realize that CPM performance in its implementation setting will probably not be *exactly* the same as in validation studies (there will likely always be some differences between the populations used for validation and implementation that are not captured by the CPM). In particular, AI-based CPMs may be sensitive to small differences between the validation and implementation populations. It may therefore be useful to consider whether another local validation is desirable before implementation [21]. If performance is then deemed suboptimal, it can be decided to update the CPM in the local setting. This updating can be done to different extents, ranging from only recalibration (adapting the intercept or slope of a regression model) to rerunning the full model development pipeline including hyperparameter tuning and feature selection. Note that new internal validations (and possibly external validations as well) are required after updating a CPM [21,22].

Second, clinical practice may change over time, and this may render any previous (internal or external) validations outdated [20,23]. For example, when a new high-sensitive D-dimer blood assay becomes available, this may influence the effectiveness of the combined predictive accuracy of the Wells score with the assay [24]. Additional validation could then be considered to assess performance in the new context. Again, any adaptations made to a CPM or its context of use will require new validations [11,21].

Note that the extent to which CPM adaptations during updating are desirable depends on the need for robustness of CPM performance over multiple settings. Some CPMs are intended to perform well specifically in a local setting (eg, UrinCheck), whereas other CPMs are intended to perform reasonably well in many settings (eg, the Wells score). CPMs intended for local use can often be designed to provide the most accurate predictions for the specific setting. For example, UrinCheck uses some highly specific predictors that are not routinely measured in all hospitals. This is different for other CPMs that are intended to be used in many settings, and that are designed to have good transportability at the cost of not providing the most accurate predictions feasible in a local setting. Updating should thus correspond to the intended use: some CPMs are updated to closely fit the local population, whereas other CPMs are updated in manners that aim to retain transportability to

Box 3 Summary of the 3 misconceptions and clarifications

"External validation provides stronger evidence of model performance than internal validation"
- Internal validation and external validation serve different purposes; either or both may be recommended depending on the intended model use

"External validation proves the generalizability of model performance"
- External validity is not a license to use a model in any setting
- External validation studies need to reflect the target population and clinical context of intended use

"External validation is a single-step process before implementation in clinical practice"
- It may be desirable to update the model for optimal performance in the local setting of intended use, which requires at least internal validations afterward
- Often model monitoring is needed to account for changes over time and new validations may be needed after adaptations to the model or its context of use

multiple populations. For example, if the performance of UrinCheck has deteriorated over time, CPM updating may entail retraining UrinCheck in the new situation and evaluating model performance by internal validation. On the other hand, when an update with an age-adjusted D-dimer cutoff value was proposed to improve the poor performance of the Wells score in older people, multiple external validation studies were performed to assess the performance of the Wells score with this new D-dimer cutoff in several populations [14,25].

Thus, external validation is *no* one-time procedure that renders the CPM "good to go" for implementation in clinical practice. CPMs may benefit from updating to the local setting of use, and from continuous monitoring and updating to account for temporal shifts in performance. The extent of changes made to a CPM in the process of updating depends on the desired balance between local performance and transportability to multiple settings.

3. Conclusion

Confusion about the concept of external validation may impede the successful implementation of many developed CPMs in clinical practice. In this article, we have discussed 3 common misconceptions about external validation to clarify when external validations are needed, and how they can be designed and interpreted (Box 3). In short, the need for and extent of external validations depend on the intended use of the CPM, and validation methods should be tailored toward that. Misconceptions can be overcome if authors explicitly report on the intended use of the CPM, the rationale for performing an external validation study, and the rationale for its design (eg, transportability with regard to specific aspects of the study population). The Figure translates the insights from this paper into considerations that stakeholders could reflect on at different stages of the validation phase in a CPM lifecycle. First, stakeholders may consider whether external validations are indeed needed. If so, it would be useful to reflect on the representativeness of existing validation studies for the intended target population of use, and potential hiatus in transportability assessments that still need to be evaluated in new external validations. Lastly, it may be worth considering CPM updating, particularly for very locally implemented CPMs or after performance shifts over time. These considerations could help stakeholders to consider, design, interpret, and appreciate external validation studies, facilitating choices for CPM implementation in clinical practice.

CRediT authorship contribution statement

Hannah M. la Roi-Teeuw: Writing – review & editing, Writing – original draft, Visualization, Project administration, Conceptualization. **Florien S. van Royen:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Anne de Hond:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Anum Zahra:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Sjoerd de Vries:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Richard Bartels:** Writing – review & editing, Conceptualization. **Alex J. Carriero:** Writing – review & editing. **Sander van Doorn:** Writing – review & editing, Conceptualization. **Zoë S. Dunias:** Writing – review & editing. **Ilse Kant:** Writing – review & editing, Conceptualization. **Tuur Leeuwenberg:** Writing – review & editing, Conceptualization. **Ruben Peters:** Writing – review & editing, Conceptualization. **Laura Veerhoek:** Writing – review & editing, Conceptualization. **Maarten van Smeden:** Writing – review & editing, Conceptualization. **Kim Luijken:** Writing – review & editing, Visualization, Supervision, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

There are no competing interests for any author.

Acknowledgments

The authors of this article have various backgrounds in clinical care, epidemiology, classical statistics, data science and/or artificial intelligence, and participate in a special interest group on clinical prediction modeling at the University Medical Center, Utrecht, The Netherlands. Discussions in the special interest group provided insights on external validation that formed the basis for the articles' content. KL initiated the idea to capture these insights in co-writing sessions for the current article. HMlRT let the project. HMlRT, FvR, SdV, RB, AdH, SvD, IK, TL, RP, LV, MvS and KL were involved in the design of the paper's outline. A writing team consisting of HMlRT, FvR, SdV, AdH and AZ discussed the outline, supervised by KL, and wrote and revised the first drafts of the manuscript. AZ created the Figure, with input from the other writing team members. All authors participated in revision of the manuscript and agreed to its final version.

References

[1] Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ 2019;365:1737.
[2] de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5:1−13.
[3] Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. Thromb Haemost 2000;83:416−20.
[4] de Vries S, ten Doesschate T, Totté JEE, Heutz JW, Loeffen YGT, Oosterheert JJ, et al. A semi-supervised decision support system to facilitate antibiotic stewardship for urinary tract infections. Comput Biol Med 2022;146:105621.
[5] Harrell F. Multivariable modeling strategies. In: Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer; 2015:63−102.
[6] van Royen FS, Moons KGM, Geersing GJ, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. Eur Respir J 2022;60:2200250.
[7] Hameed M, Yeung J, Boone D, Mallett S, Halligan S. Meta-research: how many diagnostic or prognostic models published in radiological journals are evaluated externally? Eur Radiol 2023;1:1−10.
[8] Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop 2021;92:385−93.
[9] Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25−34.
[10] Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res 2022;6:1−6.
[11] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691−8.
[12] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245.
[13] Van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J 2022;43:2921−30.
[14] Geersing GJ, Takada T, Klok FA, Büller HR, Courtney DM, Freund Y, et al. Ruling out pulmonary embolism across different healthcare settings: a systematic review and individual patient data meta-analysis. PLoS Med 2022;19:e1003905.
[15] de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. NPJ Digit Med 2023;6:1−3.
[16] Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2:e489−92.
[17] Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol 2002;55:1201−6.
[18] Lam HR, Chow S, Taylor K, Chow R, Lam H, Bonin K, et al. Challenges of conducting research in long-term care facilities: a systematic review. BMC Geriatr 2018;18:1−11.
[19] Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. J Am Med Inform Assoc 2019;26:1645.
[20] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023;21:1−8.
[21] Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol 2021;22:316.
[22] Dankowski T, Ziegler A. Calibrating random forests for probability estimation. Stat Med 2016;35:3949.
[23] van Royen FS, Asselbergs FW, Alfonso F, Vardas P, van Smeden M. Five critical quality criteria for artificial intelligence-based prediction models. Eur Heart J 2023;44:4831−4.
[24] Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med 2019;38:3444−59.
[25] Schouten HJ, Geersing GJ, Oudega R, van Delden JJM, Moons KGM, Koek HL. Accuracy of the Wells clinical prediction rule for pulmonary embolism in older ambulatory adults. J Am Geriatr Soc 2014;62:2136−41.