Validation, updating and impact of clinical prediction rules: A review
D.B. Toll, K.J.M. Janssen, Y. Vergouwe, K.G.M. Moons*
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
Accepted 14 April 2008

Abstract
Objective: To provide an overview of the research steps that need to follow the development of diagnostic or prognostic prediction rules. These steps include validity assessment, updating (if necessary), and impact assessment of clinical prediction rules.

Study Design and Setting: Narrative review covering methodological and empirical prediction studies from primary and secondary care.

Results: In general, three types of validation of previously developed prediction rules can be distinguished: temporal, geographical, and domain validations. In case of poor validation, the validation data can be used to update or adjust the previously developed prediction rule to the new circumstances. These update methods differ in extensiveness, with the easiest method a change in model intercept to the outcome occurrence at hand. Prediction rules—with or without updating—showing good performance in (various) validation studies may subsequently be subjected to an impact study, to demonstrate whether they change physicians’ decisions, improve clinically relevant process parameters, patient outcome, or reduce costs. Finally, whether a prediction rule is implemented successfully in clinical practice depends on several potential barriers to the use of the rule.

Conclusion: The development of a diagnostic or prognostic prediction rule is just a first step. We reviewed important aspects of the subsequent steps in prediction research. © 2008 Elsevier Inc. All rights reserved.

Keywords: Prediction rule; Validation; Generalizability; Updating; Impact analysis; Implementation

1. Introduction

Prediction rules or prediction models, often also referred to as decision rules or risk scores, combine multiple predictors, such as patient characteristics, test results, and other disease characteristics, to estimate the probability that a certain outcome is present (diagnosis) in an individual or will occur (prognosis). They intend to aid the physician in making medical decisions and in informing patients. Table 1 shows an example of a prediction rule.

In multivariable prediction research, the literature often distinguishes three phases: (1) development of the prediction rule; (2) external validation of the prediction rule (further referred to as “validation”), that is, testing the rule’s accuracy and thus generalizability in data that was not used for the development of the rule, and subsequent updating if validity is disappointing; and (3) studying the clinical impact of a rule on physician’s behavior and patient outcome (Table 2) [1—5]. A fourth phase of prediction research may be the actual implementation in daily practice of prediction rules, which endured the first three phases [4]. A quick Medline-search using a suggested search strategy [6] demonstrated that the number of scientific articles discussing prediction rules has more than doubled in the last decade; 6,744 published articles in 1995 compared to 15,662 in 2005. A striking fact is that this mainly includes papers concerning the development of prediction rules. A relatively small number regards the validation of rules and there are hardly any publications showing whether an implemented rule has impact on physician’s behavior or patient outcome [3,4].

Lack of validation and impact studies is unfortunate, because accurate predictions—commonly expressed in good calibration (agreement between predicted probabilities and observed outcome frequencies) and good discrimination (ability to distinguish between patients with and without the outcome)—in the patients that were used to develop a rule are no guarantee for good predictions in new patients, let alone for their use by physicians [1,3,4,7,8]. In fact, most prediction rules commonly show a reduced accuracy when validated in new patients [1,3,4,7,8]. There may be two main reasons for this: (1) the rule was inadequately developed and (2) there were (major) differences between the derivation and validation population.
Many guidelines regarding the development of prediction rules have been published, including the number of potential predictors in relation to the number of patients, methods for predictor selection, how to assign the weights per predictor, how to shrink the regression coefficients to prevent overfitting, and how to estimate the rule’s potential for optimism using so-called internal validation techniques such as bootstrapping [1,2,7–14].

Compared to the literature on the development of prediction rules, the methodology for validation and studying the impact of prediction rules is underappreciated [1,4,8]. This paper provides a short overview of the types of validation studies, of possible methods to improve or update a previously developed rule in case of disappointing accuracy in a validation study, and of important aspects of impact studies and implementation of prediction rules. We focus on prediction rules developed by logistic regression analysis, but the issues largely apply to prediction rules developed by other methods such as Cox proportional hazard analysis or neural networks. The methodology applies both to diagnostic and prognostic prediction rules and is illustrated with examples from diagnostic and prognostic research.

2. Examples of disappointing accuracy of prediction rules

Even when internal validation techniques are applied to correct for overfitting and optimism, the accuracy of prediction rules can be substantially lower in new patients compared to the accuracy found in the patients of the development population. For example, the generalizability of an internally validated prediction rule for diagnosing a serious bacterial infection in children presenting with fever without apparent source was disappointing [15]. In the development study, the area under the receiver operating characteristic curve—after adjustment for optimism—was 0.76 (95% confidence interval: 0.66–0.86). However, when applied to new patients obtained from another hospital in a later period using the same inclusion and exclusion criteria, the area under the receiver operating characteristic curve dropped to 0.57 (95% confidence interval: 0.47–0.67). The authors concluded that this could partly have been caused by flaws in the development of the rule, notably too few patients in relation to the number of predictors tested, but also that internal validation and correction for optimism do not always prevent a decreased accuracy in future patients [15].

Another example of poor generalizability regards the European System for Cardiac Operative Risk Evaluation (EuroSCORE), a prognostic prediction rule that was developed in 128 centers in eight European states to predict 30-day mortality in patients who underwent cardiac surgery [16,17]. Validation studies showed good results in European, North American, and Japanese populations [16,18–23]. Yap et al. [24] tested the generalizability of the EuroSCORE in 8,331 cardiac surgery patients from six Australian institutions and found that predictions were poorly calibrated for Australian patients. According to the authors, reasons for this finding are unclear and likely to be multifactorial, such as a different health care system, different indications for cardiac surgery, and a different prevalence of comorbid conditions in Australia compared to Europe. Also, the prediction rule could be “out of date” [24,25]. The rule was developed with data of patients who were operated more than 10 years prior to the patients in the Australian validation study. The surgical procedure has indeed changed over time, potentially leading to different outcomes [24,25]. This change was not reflected by the other validation studies [16,18–23].

2.1. Common differences between a development and a validation population

As described, disappointing generalizability can be explained by differences in the development and validation population. We may largely identify three possible differences. First, the definitions of predictors and the outcome variable and the measurement methods may be different [1,2,8,11]. Prediction rules that contain unclear defined predictors or predictors which measurement or interpretation is likely to show a reduced predictive
strength when applied to new patients. For example, the diagnostic prediction rule of Table 1 contains the rather atypical predictor gender, but also the presence of “atypical convulsions” [26]. The latter may be defined differently by physicians, which may compromise the generalizability of the rule. It is advised to determine the interobserver variability of potential predictors and to include only those predictors in the final prediction rule that show good reliability [2,5]. Improvement in measurement techniques for predictors may also affect the predictive strength of a predictor. For example, the Magnetic Resonance Imaging technique is developing rapidly over time, which results in improved image quality. Consequently, the diagnostic or prognostic information of Magnetic Resonance Imaging probably also improves over time and influences the accuracy of prediction rules that include Magnetic Resonance Imaging information.

Second, the group of patients used for the development of a prediction rule may be different from the group of patients used for validation. This is also called difference in “case mix” [1,27]. For example, differences in indication for cardiac surgery and differences in comorbidity were considered as one of the causes of the poor calibration of the prognostic EuroSCORE in Australian patients. Both discrimination and calibration of a rule can be affected by differences in case mix. For example, a validation population may only include elderly (e.g., defined as age ≥65 years), whereas in the development population individuals’ age ranged from 18 to 85 years. If age is a predictor in the rule, then discrimination between presence and absence of the outcome in the more homogeneous validation population is more difficult than in the more heterogeneous development population. Further, a validation population may, for example, contain relatively more males than the development population. If male gender increases the probability of the outcome but gender was not included in the rule (missed predictor), then the predicted probabilities by the rule will be underestimated in the validation population (reduced calibration) [27].

Third, validation studies commonly include fewer individuals than development studies. Accordingly, both populations may seem different, which is notably due to random variation [12,28]. The required size of a validation study depends on the hypotheses tested. For prediction rules that predict dichotomous outcomes, it has been suggested that the validation sample should contain at least 100 events and 100 nonevents to detect substantial changes in accuracy (for example, a 0.1 change in c-statistic) with 80% power [29].

2.2. Type of validation studies

It has repeatedly been suggested that a validation study should consist of an adequate sample of “different but related patients” compared to the development study population [1,4,8]. Relatedness is at least defined as “patients suspected of the same disease” for a diagnostic rule, and as “patients at risk of the same event” for a prognostic rule.

In hierarchy of increasingly stringent validation strategies, we largely distinguish temporal, geographical, and domain validations [1,3,4,8]. In general, the potential for differences between the development and validation population is smallest in a temporal validation study, and largest in a domain validation study (Table 3). Consequently, confirmative results in a domain validation study are considered to provide the strongest evidence that the prediction rule can be generalized to new patients, whereas the generalizability of a prediction rule that has shown confirmative results in a temporal validation study may still be restricted.

When various published prediction rules for the same outcome are available from different databases, these rules can be mutually validated in these databases which sometimes have even been prepared to undertake such a cross-validation. It is also possible to validate the available prediction rules in one comprehensive database including data from various time periods, geographical areas, and even clinical domains [30].

2.2.1. Temporal validation

Temporal validation tests the generalizability of a prediction rule “over time.” In a temporal validation study, the prediction rule is typically tested by the same physicians or investigators as in the development study, in the same institution(s), and in similar patients, for example, using the same eligibility criteria resulting in small variation in case mix (Table 3) [3,4,8]. Hence, this type of validation is usually successful for thoughtfully developed prediction rules.

---

Table 2
Consecutive phases in multivariable prediction research, diagnostic or prognostic

<table>
<thead>
<tr>
<th>Phase</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Development</td>
<td>Development of a multivariable (diagnostic or prognostic) prediction rule, including identification of important predictors, assigning the relative weights to each predictor, estimating the rule’s predictive accuracy, estimating the rule’s potential for optimism using so-called internal validation techniques, and—if necessary—adjusting the rule for overfitting.</td>
</tr>
<tr>
<td>2. Validation and updating</td>
<td>Testing the accuracy of the prediction rule in patients that were not included in the development study. Temporal, geographical, and domain validations can be distinguished. If necessary, the prediction rule can be updated, by combining the information captured in the rule (development study) and the data of the new patients (validation study).</td>
</tr>
<tr>
<td>3. Impact</td>
<td>Determining whether a (validated) prediction rule is used by physicians, changes therapeutic decisions, improves clinically relevant process parameters, improves patient outcome or reduces costs.</td>
</tr>
<tr>
<td>4. Implementation</td>
<td>Actual dissemination of the diagnostic or prognostic prediction rule in daily practice to guide physicians with their patient management.</td>
</tr>
</tbody>
</table>
Table 3
Potential differences between the development and validation population and the influence on generalizability of the prediction rule

<table>
<thead>
<tr>
<th>Differences in</th>
<th>Temporal validation: to test the generalizability of a prediction rule “over time” in similar patients as in the development study (same inclusion and exclusion criteria) in the same hospitals or institutions</th>
<th>Geographical validation: to test the generalizability of a prediction rule in similar patients as in the development study (same inclusion and exclusion criteria) in hospitals or institutions of another geographical area</th>
<th>Domain validation: To test the generalizability of a prediction rule across different domains, which may contain other patient (sub)groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation of predictors and outcome</td>
<td>±: Often same physicians as in the development study who are thus experienced in obtaining the predictors and outcome; differences in interpretation of predictors is unlikely</td>
<td>+: Other physicians as in the development study, who may define subjective predictors (and the outcome) differently; differences in interpretation of predictors may occur</td>
<td>+: Other physicians as in the development study, who may define subjective predictors (and the outcome) differently; differences in interpretation of predictors may occur</td>
</tr>
<tr>
<td>Used measurements for predictors and outcome</td>
<td>±±: Same measurements used, unless measurements have been replaced (time related)</td>
<td>+: Other measurements may be used (&quot;institution dependent&quot;), plus potential for time-related changes in measurements</td>
<td>+: Other measurements may be used (&quot;institution dependent&quot;), plus potential for time-related changes in measurements</td>
</tr>
<tr>
<td>Case mix</td>
<td>±: Patient populations are similar; (random) variation due to a commonly small sample size of the validation population is possible</td>
<td>±± (Subtle) differences in case mix possible, beyond possible (random) variation due to a commonly small sample size of the validation population</td>
<td>++: Differences in case mix are (very) likely, beyond possible (random) variation due to a commonly small sample size of the validation population</td>
</tr>
</tbody>
</table>

--; Weak; ±±: Possible; ±: Probable; ++: Likely.

However, improvements in medical techniques may still affect the predictive accuracy of a prediction rule, as in the example of the EuroSCORE. Confirmative results in (multiple) temporal validation studies indicate that clinicians may cautiously use the prediction rule in their future patients who are similar to the development and validation population. But validation in varied study sites is still necessary before the rule can be implemented in other (geographical) locations or other patient domains (see below) [1,4,8].

2.2.2. Geographical validation

Geographical validation studies typically test the generalizability of a prediction rule in a patient population that is similarly defined as the development population (as is the case in temporal validation studies), though in hospitals or institutions of other geographical areas [1,8]. “Other geographical areas” can be within one country, across similar countries (e.g., western to western or nonwestern to nonwestern countries), and across nonsimilar countries (e.g., western to nonwestern or vice versa). Understandably, the less similar the development and validation locations are, the more potential for differences in “interpretation of predictors and outcome,” “measurements used,” and “case mix,” and thus the more potential for disappointing generalizability (Table 3).

Physicians participating in a geographical validation study may be less experienced using the prediction rule, which may influence the accuracy of the rule, due to predictors sensitive to subjectivity. Further, measurement of
predictors and the outcome can be performed with different methods than in the development study, also potentially affecting the rule’s accuracy. For example, prediction rules to safely exclude the diagnosis deep vein thrombosis (DVT) contain items from patient history, physical examination, and the result of a D-dimer test [31,32]. Many different D-dimer assays are available, each with different diagnostic accuracy measures. Sensitivities of D-dimer tests for diagnosing DVT may vary between 48% and 100%, and specificities from 5% to 100% [33]. This variation in the accuracy of different D-dimer tests will obviously be reflected in the accuracy of the diagnostic rules that include D-dimer tests. Further, if different cutoff values are used to dichotomize a continuous variable to define a positive vs. negative result, the predictive accuracy of the same variable may be different across studies. For instance, if an abnormal D-dimer concentration has been defined as higher than 500 ng/mL in the development sample and as higher than 1,000 ng/mL in the validation sample, the rule will likely perform differently in the validation sample.

Case-mix differences in a geographical validation study can be subtle. For example, a rule containing “antibiotic use in previous month” as a predictor to estimate the probability of 30-day hospitalization or death from lower respiratory tract infection in elderly patients [34] may show decreased accuracy in another country, not because the predictor was not well defined or liable to subjectivity, but because the indication for prescribing antibiotics, and thus the characteristics of antibiotic receivers, may vary between countries.

Geographical validation studies can be done prospectively but also retrospectively using previously collected data that contain all predictors and the outcome(s) of interest.

2.2.3. Domain validation

Perhaps the broadest form of validation is to test the generalizability of a prediction rule across different domains, such as patients from a different setting (primary, secondary, or tertiary care), inpatients versus outpatients, patients of different age categories (e.g., adults vs. adolescents or children), of a different gender, and perhaps from a different type of hospital (academic vs. general hospital). Obviously, the case mix of a patient population of a new domain will differ from the development population (Table 3), which is usually reflected in differences in the distribution of the predictor values and in the ranges of predictor values.

For example, the case mix of primary care and secondary care patients is often clearly different. Primary care physicians always selectively refer patients to specialists. These referred patients commonly have relatively more severe signs or symptoms or have a relatively more developed disease stage [8,35]. Consequently, secondary care patients commonly have a narrower range of the (more severe) predictor values than primary care patients. To some extent, a secondary care population can be considered as a subdomain of the primary care population. Hence, in contrast, validating a prediction rule developed from secondary care in a more heterogeneous primary care population actually concerns the estimation of the rule’s ability for extrapolation. Extrapolation of prediction rules developed in secondary care to primary care patients often results in a decreased accuracy [8,35]. For example, it has been shown that a prediction rule for safely excluding the diagnosis DVT developed in secondary care [32,36] showed disappointing accuracy in primary care: 0.9% of the secondary care patients had DVT while the diagnosis was ruled out according to the rule, whereas this proportion was increased to 2.9% in primary care patients [37].

We note, however, that certain inclusion criteria in development studies may have been chosen for practical reasons only, and may not compromise the generalization of a prediction rule derived from such studies. For example, the Ottawa ankle rule for safely excluding fractures without additional X-ray testing [38] was developed on patients aged 18 years or older. One could question whether this rule—which does not include age as a predictor—is accurate when applied to children or adolescents. If the relative weights (odds ratios) of the predictors in the rule are independent of age, extrapolation of the rule to adolescents can be as successful as in the development population. A recent review concerning the extrapolation of the Ottawa ankle rule to children or adolescents indeed concluded that “a small” percentage (1.4%) of patients that are excluded from receiving X-ray evaluation based on the Ottawa ankle rule will actually have a fracture [39], compared to 0% in the development study (in the development study, the score threshold was specifically chosen to achieve a 100% negative predictive value).

Like geographical validation, domain validation can be done prospectively but also retrospectively using previously collected data.

3. Updating prediction rules

When a validation study shows disappointing results, researchers are often tempted to reject the rule and directly pursue to develop new rules with the data of the validation population only. However, although the original prediction rules usually have been developed with large data sets, validation studies are frequently conducted with much smaller patient samples. The redeveloped rules are thus also based on smaller samples. Furthermore, it would lead to many prediction rules for the same outcome, obviously creating impractical situations, as physicians have to decide on which rule to use. For example, there are over 60 published rules to predict outcome after breast cancer [40]. Moreover, when every new patient sample would lead to a new prediction rule, prior information that is captured in previous studies and prediction rules would be neglected. This is counterintuitive to the intention that scientific inferences
should be based on data of as many patients as possible. This principle of using prior knowledge from previous studies has been recognized and used in etiologic and intervention research, for example, in the realm of (cumulative) meta-analyses.

A logical alternative to redeveloping prediction rules in each new patient sample, is to update existing prediction rules with the data of the new patients in the validation study. As a result, updated rules combine the prior information that is captured in the original rules with the information of the new patients of the validation population [41–44]. Hence, updated rules are adjusted to the characteristics of the new patients, and likely show improved generalizability.

Several updating methods have been proposed in the literature [41–44]. The methods vary in extensiveness, which is reflected by the number of parameters that is adjusted or re-estimated (Table 4). We will briefly describe these methods and refer to the literature for a more profound description [41–44].

In many situations, as described before, differences in outcome incidence are found between the development data and the validation data. For example, in a primary care setting one can validate a secondary care rule that predicts the presence or absence of DVT. Due to the higher prevalence of DVT in secondary care [35], the calibration of the rule in primary care patients may be poor as a result of systematically too high predicted probabilities. By adjusting only the intercept of the original prediction rule for the patients in the primary care setting, the poor calibration can be improved [43,44]. This method is by far the simplest updating method as only one parameter of the original rule, that is, the intercept, is adjusted (Table 4, method 1).

Another updating method is called “logistic recalibration” and can be used when the regression coefficients (relative weights that represent the predictive strength) of the predictors in the prediction rule are overfitted in the development study [43,44]. This typically results when too many predictors were considered in a too small dataset [7,13]. In the lower range, the predicted probabilities in the new patients are usually too low, whereas in the higher range they are too high. When overfitting was not adequately prevented or adjusted during development of the rule, all regression coefficients can still be adjusted with a single correction factor that is easily estimated from the data of the new patients in the validation set (Table 4, method 2).

Although calibration can indeed be improved by these first two methods, discrimination (area under the receiver operating characteristic curve) will remain unchanged, as the relative ranking of the predicted probabilities remain the same. To improve the discrimination of a rule in new patients, more rigorous adjustments need to be made to the prediction rule, also called model revisions. We will briefly explain four revision methods that can improve both the discrimination and the calibration of a prediction rule when a rule is validated in new patients.

First, the strength of one or more predictors can be different in the validation population compared to the development population, whereas the relative sizes of the other regression coefficients to each other are correct. The regression coefficients that differ can be re-estimated from the validation data (Table 4, method 3). For example, we discussed a prediction rule with the antibiotic use as a predictor (Section 2.2.2). When this rule is applied in a population or setting with a different antibiotic prescription strategy, the strength of this particular predictor may be different, whereas the strength of the other predictors in the rule not necessarily changes.

Also, when potential predictors that may have predictive value were not included in the original rule, one can test whether these have added predictive value in the validation data (Table 4, method 4). For example, when a prediction rule is validated over time, and a new test has become available, the new test may have added predictive value in the rule.

Finally, when the previous described updating methods cannot improve the accuracy of the rule, and the strengths of all predictor are expected to be different in the new patients, the intercept and the regression coefficients of all predictors can be re-estimated with the data of the new patients (Table 4, method 5). If necessary, additional predictors can be considered as well (Table 4, method 6). These two methods are the most rigorous updating methods, as the intercept, regression coefficients and possibly also additional predictors are all re-estimated from the validation set. Both methods will probably be most applicable to domain validation, as these are typical situations in which the strength of predictors may differ between the two populations. Note that a disadvantage of these rigorous updating methods is that the rule is redeveloped on the data of the validation set only and that the prior information in the original rule is neglected, as we discussed above.

With all above described methods, the updated rules are adjusted to the circumstances of the validation population. However, we recommend that updated prediction rules, just like newly developed rules, still need to be tested on their generalizability and impact before they can be applied in daily practice. Note that for all updating methods, data of the new patients is needed. When this data is not available, but one knows the incidence of the outcome and the mean value of the predictors in the new population, the rule can be adjusted by a simple adjustment of the prediction rule [45,46].

4. Impact analysis

To ascertain whether a validated diagnostic or prognostic prediction rule will actually be used by physicians, will change or direct physicians’ decisions, and will improve clinically relevant process parameters (such as number of bed days, length of hospital stay, or time to diagnosis), patient outcomes, or reduces costs, an impact study or impact analysis should be performed [3,4]. In the ideal design of an
impact study, physicians or care units are randomized to either the index group—which is “exposed” to the use of the prediction rule—or to the control group using “care or clinical judgment as usual”: a cluster randomized trial [3,47]. There have been some nice examples of cluster-randomized studies of the impact of prediction rules [48–50]. Randomization of patients instead of physicians—such that a physician randomly uses the prediction rule or applies “usual care”—is not advised. Learning effects will lead to a reduced contrast between the two study groups, resulting in a diluted measured impact of the rule. Moreover, randomizing centers (requiring a multicenter study) instead of physicians within a single center may prevent the risk of contamination—i.e., exchange of experiences and information by physicians between the two study groups—also leading to reduced contrast and dilution of the rule’s effect. Patients are followed up to determine the impact of the prediction rule on clinically relevant process parameters, patient outcome, and on cost-effectiveness. Randomizing physicians to different diagnostic or prognostic strategies—e.g., with and without exposure or use of a diagnostic or prognostic prediction rule—can only lead to differences in patient outcome if it leads to actual differences in treatment decisions. Follow up is not required for studying the influence of a prediction rule on decision-making behavior: a cross-sectional, randomized study then suffices. An alternative design to determine the impact of a prediction rule is a before—after study within the same physicians or care units [51], although temporal changes may compromise the validity of this design.

Although an impact analysis is the method par excellence to study the real effect of a prediction rule in practice, only a limited number of impact analyses have been performed [4]. One of these few studies regarded the impact of a prediction rule aimed at improving the effectiveness of treatment of patients presenting with community-acquired pneumonia to the emergency department, measured by the health-related quality of life and the number of bed days per patient [49]. Hospitals were randomly assigned to implementation of the prediction rule or conventional care, by using a computer that stratified for type of institution (teaching or community hospital) and average length of stay. Physicians were instructed to use the prediction rule as a guide only; the rule did not supersede clinical judgment. An educational plan was designed to reinforce compliance with the use of the prediction rule. The authors concluded that the prediction rule did not improve the health-related quality of life of the patients but reduced the number of bed days per patient managed. This effect was never revealed if no impact study had been conducted.

5. Implementation of prediction rules

When a rule has frequently been proven to be accurate in diverse populations, the more likely it is that the prediction rule can be successfully applied in practice [1,4,8]. Yet, there are still reasons why the rule is not as successful in daily practice.

First, physicians may feel that their often implicit estimation of a particular predicted probability is at least as good as the probability calculated with a prediction rule, and may therefore not use or follow the rule’s predictions [3]. Or, the physicians’ estimation of a probability may even have proven to outperform the discrimination of a prediction rule. In a recent review, Sinuff et al. compared the discrimination of physicians’ estimations with prediction rules in predicting the mortality of critically ill patients considered for intensive care unit (ICU) admission [52]. They concluded that ICU physicians more accurately discriminate between survivors and nonsurvivors in the first 24 h of ICU admission than prediction rules do. It may thus be important to compare physicians’ predictions with those of prediction rules, preferably already during the development phase of a prediction rule (requiring obviously a prospective design), but certainly in validation and impact studies. If physicians’ predictions of probabilities have proven to outperform the probabilities provided by a rule, the rule will likely not be used in practice. In contrast, results in favor for the prediction rule can be used to convince physicians of using the rule when properly validated.

Second, prediction rules must have face validity; physicians must accept the logic, as well as the science of the rule. Clinical prediction rules that do not have face validity may not be applied in practice, even when effective [53].

Third, prediction rules may not be used because they are not user-friendly [3,54]. The user-friendliness should be taken into account when developing the rule. A variable should only be considered as a potential predictor if obtaining this variable is also feasible in daily practice of the type of patients under study (not too time-consuming or costly). The user-friendliness of a prediction rule also depends on the way a prediction rule is presented; the original regression formula (e.g., as in the legend of Table 1) is the most exact and accurate form, but may involve cumbersome calculations requiring a calculator or computer. Although sumscores, risk stratification charts, or nomograms may be less precise, they certainly are more user-friendly. With the introduction of electronic patient records, the use of regression formulas in daily practice will become much easier.

Finally, practical barriers may exist to act on the results of the prediction rule. For instance, when using a diagnostic prediction rule aiming to determine whether subsequent testing is necessary, such as the Ottawa ankle rules, physicians may be concerned about protecting themselves against litigation. Hence, they may still refer their patients for additional testing, whereas at the same time the prediction rule indicates that referral was not necessary [3]. Brehaut et al. [55] conducted a postal survey among 399 randomly selected physicians to examine whether physicians used the Ottawa ankle rules [38] for diagnosing ankle
fractures. Most physicians (90%) reported to use the Ottawa ankle rules always or most of the time in appropriate circumstances, whereas only 42% actually based their decisions to order radiography primarily on the rule [55]. The same authors assessed why some prediction rules become widely used and others do not [56], taking the Canadian Cervical Spine Rule [57] as an example. They showed that older physicians and part-time working physicians were less likely to be familiar with the rule. The best predictors whether a rule would be used in practice were the familiarity acquired during training, the confidence in the usefulness of the rule, and the user-friendliness of the rule [56].

6. Final comments

We have given an overview of types of validation studies, of methods to improve or update a previously developed diagnostic or prognostic prediction rule in case of disappointing accuracy in a validation study, and of aspects of impact studies and the implementation of prediction rules. A validated, and if necessary updated, rule may cautiously be applied in new patients that are similar to the patients in the development and validation populations. However, when the user has reasons to believe that the rule may perform differently in the new patients, data of the new patients should first be collected to test the accuracy, and preferably impact of the rule, before the rule is applied in daily clinical practice. Any rule may perform slightly different in a new patient sample due to sampling variation. In that situation, the rule does not need to be updated. The questions remains: when has a rule been sufficiently validated and updated? So far, this particular methodological area of prediction research has not been explored. Future research should address the question how many validation studies and what type of adjustments are needed before it is justified to implement a prediction rule into clinical practice. Further, the performance of prediction rules may worsen over time, as they can become “out of date” (as mentioned in Section 2). Especially in rapidly developing medical disciplines, physicians should first verify whether the medical situation nowadays is still comparable to the situation during the development of the prediction rule, before applying the rule in their practice. Regular “postmarketing surveillance” to evaluate whether the accuracy of the prediction rules holds over time, might be something to consider. Such repeated validations could indicate whether a rule should no longer be used, or at least updated to the new circumstances. The question of course arises how often such repeated validation studies should be undertaken, which is a topic for further research.

Another subject of prediction research that may need more focus in the future is the methodology for systematic reviews and perhaps even meta-analyses of prediction rules for the same outcome [58,59]. It will be a challenge to define how regression coefficients of prediction rules can be combined, and how to properly address publication bias; as prediction rules with good results are more likely to be published than rules with moderate results. Although the methodology for meta-analyses has been extensively described for etiologic and intervention studies, to our knowledge no research has been conducted for meta-analysis to combine several prediction rules.

Last, the potential gain in predictive accuracy and generalizability of a prediction rule developed on combined datasets with individual patient data from various studies on the same outcome (so-called individual patient studies) is a research area that needs more attention [60].

Our purpose was to stress the importance of testing the generalizability and impact of prediction rules, and outline the methods of such research. The relevance and importance of validating and testing the impact of prediction rules on physician’s behavior, and patients’ outcome, has repeatedly been emphasized in the literature. Unfortunately, only a relatively small number of rules are validated, and hardly any study questions whether an implemented rule can change patient outcome. An increased focus on validation and impact studies will likely improve the application of valid prediction rules in daily clinical practice.

Acknowledgments

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZonMw 016.046.360; ZonMw 945-04-009).

References


