Poor Agreement among Prothrombin Time International Normalized Ratio Methods: Comparison of Seven Commercial Reagents

Juha Horsti,* Helena Uppa, and Juhani A Vilpo

Background: Prothrombin time (PT) has long been the most popular test for monitoring oral anticoagulation therapy. The International Normalized Ratio (INR) was introduced to overcome the problem of marked variation in PT results among laboratories and the various recommendations for patient care. According to this principle, all reagents should be calibrated to give identical results and the same patient care globally. This is necessary for monitoring of single patients and for application of the results of anticoagulation trials and guidelines to clinical practice.

Methods: We took blood samples from 150 patients for whom oral anticoagulation had been prescribed. Plasmas were separated and PTs determined by use of seven commercial reagents and four calibrator sets. The differences in results were assessed by plotting, for each possible pair of methods, the differences in INR values for each sample against the mean INR value (Bland–Altman difference plots).

Results: Mean results differed significantly (P < 0.001) for 17 of 21 possible paired comparisons of methods. Only two pairs of methods produced very similar results when assessed for problems of substantial differences in INR values; a significant, systematic increase in the difference with INR; and a significant systematic increase in the variation in difference with increasing INR values.

Conclusions: The agreement among several (and perhaps most) commercial INR methods is poor. The failure of current calibration strategies may severely compromise both the monitoring of individual patients and the application of oral anticoagulation guidelines and trial results to clinical practice.

© 2005 American Association for Clinical Chemistry

Arterial and venous thrombotic events are responsible for more morbidity and mortality than any other condition in the developed world (1, 2). Chronic oral anticoagulation treatment with the vitamin K antagonists warfarin is effective in the prevention and treatment of thromboembolisms in various clinical situations (3–10). A clinically relevant anticoagulation effect, or hypocoagulability, can be achieved only at the cost of increasing the risk of hemorrhage. Individual responses to oral anticoagulation treatments are particularly variable and unpredictable. Frequent and adequate laboratory controls are thus required to ensure that the anticoagulation effect remains within the therapeutic range.

The prothrombin time (PT) assay was introduced to monitor chronic warfarin therapy. It primarily measures vitamin K-dependent coagulation factors II, VII, and X (6). The two most commonly used PT tests are based on Quick’s one-stage PT (11, 12) and Owren’s (13) method (combined thromboplastin reagent), which was developed to overcome the drawbacks of the Quick method; among other things, fibrinogen and factor V were added to the reagent (13). The Owren method is most widely used in the Nordic and Benelux countries as well as in Japan, whereas the Quick PT is the approach used elsewhere, accounting for ~95% of PT tests worldwide.

The results of coagulation analyses in different laboratories, even those using the same method, have been very difficult to harmonize. The comparability of PT results is essential for two reasons: (a) for the safety of the individual patient on anticoagulation therapy, and (b) to improve the applicability of anticoagulation guidelines based on

1 Nonstandard abbreviations: PT, prothrombin time; INR, International Normalized Ratio; ISI, International Sensitivity Index; and GHI, Global Hemostasis Institute.
clinical trials and expert recommendations. The harmonization of PT results is hampered by the lack of a natural primary standard. This obstacle has been circumvented by the concept of the International Normalized Ratio (INR). Prerequisite to this harmonization has been adoption by the WHO of an international reference thromboplastin preparation (13, 14): each new commercial thromboplastin is calibrated against the primary WHO reference preparation. The results are used to calculate the relative sensitivity of the unknown preparation compared with the WHO standard International Sensitivity Index (ISI) (15). The INR is calculated according to the formula: INR = [PT ratio]ISI (15).

Increasing use of the INR format has led to appreciation of its limitations (5, 16–19). These include poorer accuracy and precision when insensitive (high ISI value) thromboplastins are used, incorrect assignment of an ISI value by the manufacturer, use of a reagent–instrument combination different from that used by the manufacturer, and acceptance of an incorrect PT value for the reference plasma used for calibration and quality control (20). It is somewhat surprising that, despite the wide use of oral anticoagulation therapy and its laboratory monitoring, only limited information is available concerning agreement among results obtained with the various INR methods. Preliminary data from our laboratory point to the possibility that the Quick and Owren methods may produce different results (21–23). Moreover, some of the results may be systematically biased and clinically unacceptable (24). In the present study, we sought to establish how general these problems are by comparing the INR results for 150 blood samples and determining the INR values obtained with seven commercial INR methods, four Quick- and three Owren-based methods.

Materials and Methods

PATIENTS AND BLOOD SAMPLING

Hospital and health-center patients were eligible if a PT time test was requested for the monitoring of oral anticoagulation therapy. In our region, a “P-INR” test code is used for this purpose. Hence, the samples represented all possible phases of anticoagulation (before treatment, dose-adjustment phase, and steady-state phase) and not the chronic phase alone. Patients were chosen without conscious bias. All procedures were approved, in accordance with the Helsinki Declaration of 1975, by our institution’s responsible ethics committee.

Blood (1.8 mL) was drawn from 150 patients into citrate coagulation tubes (Vacutet cat. no. 454322, 9NC or 4544392, 9NC; Greiner Labortechnik GmbH), the great majority of them containing 0.2 mL (0.109 mol/L) of citrate solution. We received from the district health centers two or three samples in 129 mmol/L citrate tubes previously used in our laboratories. This “contamination” could not be identified, and the results are included in the final data. Sample tubes were centrifuged at 1850g for 10 min at 20 °C to separate plasma. Measurements commenced within 8 h of blood collection with minimal delay between runs.

PT DETERMINATION

The PT coagulation times were measured by use of a fully automated BCS coagulation analyzer (The Dade Behring Coagulation System). For the one-stage PT with the Quick method, 100 μL of coagulant reagent was added to 50 μL of citrated plasma. The four test reagents were as follows:

- Neoplastine CL Plus (rabbit brain thromboplastin; cat. no. 00376; lot no. 031581; Diagnostica Stago); ISI = 1.30 (no instrument mentioned); ISI values obtained with the calibration reagents listed below on the BCS analyzer: 1.21, 1.42, 1.20, and 1.39
- PT-Fibrinogen Recombinant (recombinant rabbit tissue thromboplastin; cat. no. 20005000; lot no. NO425869; Instrumentation Laboratory); ISI = 1.03 on the ACL (Instrumentation Laboratory); ISI values obtained with the calibration reagents listed below on the BCS analyzer: 0.90, 1.08, 0.94, and 1.02
- PT-Fibrinogen HS Plus (rabbit brain thromboplastin; cat. no. 08469810; lot no. NO325729; Instrumentation Laboratory); ISI = 1.13 on the ACL; ISI values obtained with the calibration reagents listed below on the BCS analyzer: 1.36, 1.25, 1.41, and 1.63
- Dade Innovin (recombinant human tissue thromboplastin; cat. no. B4212-50; lot no. 526987; Dade Behring Marburg GmbH); ISI = 0.90 on the ACL; ISI values obtained with the calibration reagents listed below on the BCS analyzer: 0.98, 1.04, 0.83, and 1.07

For the Owren PT (combined thromboplastin reagent), the coagulation reaction contained 10 μL of citrated sample plasma, 50 μL of diluent, and 150 μL of reagent. The three test reagents were as follows:

- Owren PT [rabbit brain thromboplastin; cat. no. GHI 131-10; Global Hemostasis Institute (GHI); containing 25 mmol/L CaCl2 (cat. no. GHI 155) and a diluent (Owren buffer; cat. no. GHI 150) from the GHI (lot no. C414F)]; ISI = 1.09 for optical methods; ISI values obtained with the calibration reagents below on the BCS analyzer: 1.15, 1.17, 1.24, and 1.50
- Nycotest PT [rabbit brain thromboplastin (cat. no. 1002488) and a diluent (Nycotest PT, dilution liquid, cat. no. 1002485)] from Axis-Shield (lot no. 10107353); ISI = 1.13 on the Thrombotrack; ISI values obtained with the calibration reagents below on the BCS analyzer: 1.01, 0.95, 1.08, 1.39
- SPA, 50 [thromboplastin (cat. no. 00105) and a diluent (SPA buffer cat. no. 00124)] from Diagnostica Stago; lot no. 022071; ISI = 0.98 (no instrument mentioned); ISI values obtained with the calibration reagents below on the BCS analyzer: 1.03, 0.93, 1.05, and 1.31
ISI Calibration
ISI calibration of all seven reagents was performed with each of the four available calibrators:

- Local Finnish ISI calibrators (ISI-kalibratorKit; cat. no. BL0000150, lot no. 7; Bioclin). This calibration is suitable for both the Quick and Owren PT methods
- Etaloquick (cat. no. 00496; lot no. 021964; Diagnostica Stago)
- PT Calibration Plasma Kit (cat. no. OQLZ11; lot no. OU2100H; Dade Behring Marburg GmbH)
- Calibration Set (cat. no. 1048851; lot no. 1D0100Z; Axis- Shield)

An ISI calibration reagent set contains calibrators and a normal plasma that have known INR values. The value for the PT normal plasma is the geometric mean of the PT from 20 healthy individuals. We calculated ISI values for individual reagents, using four different calibration reagent sets (three to four calibrators per reagent set). The ISI value calculation is based on the same INR equation as below and WHO recommendations, but in another form. Calibrator plasmas have known INR values, and they are traceable to the WHO international standard. Duplicate measurements (in seconds) on calibrators were made with the BCS and the formula below (15).

The ISI value is the average of calculations of different calibrators:

\[
\text{ISI} = \frac{\log(\text{INR (calibrator)})}{\log(\text{calibrator}/\text{normal})}
\]

All four ISI values for each of the seven reagents were used for analyzing and comparing patient INR values. In other words, for each sample, \(4 \times 7 = 28\) INR values were finally determined.

Analytical Imprecision of PT Determinations
We assessed the within-run imprecision of the seven PT tests, using one patient plasma sample (\(n = 10\) determinations) with an INR value in the therapeutic range, i.e., \(\sim 2.2\) INR. The respective CVs were 2.3\% for the Neoplas-tine CL Plus, 2.7\% for the PT-Fibrinogen Recombinant, 1.1\% for the PT-Fibrinogen HS Plus, 2.6\% for the Dade Innovin, 1.4\% for the Owren PT, 1.6\% for the Nycotest PT, and 1.0\% for the SPA.

INR Calculation
INR results were calculated in seconds according to the formula: \(\text{INR} = \left(\text{sample} / \text{normal}\right)^{\text{ISI}} (15)\).

Statistical Analysis
We used Friedman ANOVA and Kendall concordance analysis of the INR values for an overall assessment of the agreement among different methods; we then used the Wilcoxon matched-pairs test to compare the INR values pairwise individually.

We also examined the measuring agreement of the different PT methods, using the method described by Bland and Altman (25). For this purpose, we prepared a plot of differences between the two methods to be compared against their means (Fig. 2). Usually this method of determination of bias (mean difference or \(d\)) and standard deviation of differences (\(s\)) is a versatile tool to investigate lack of agreement between two clinical laboratory tests. We noted, however, that the mean difference is not necessarily stable at various INR values. Consequently, we introduced three additional comparison parameters: (a) the scatter ratio of INR differences; (b) the slope of the least-squares regression line correlating INR differences vs INR means; and (c) the 95\% prediction interval of the dependent variable (INR difference) for given values of the independent variables (INR mean).

The scatter ratio was the ratio of the standard deviation (scatter) of INR differences above the mean INR value of 2.5 against that below the mean INR value of 2.5. The value 2.5 was chosen arbitrarily. A scatter ratio \(>1\) indicates that the difference in INR values between the two methods increases when the INR values increase.

A slope of the regression line that is significantly different from zero (as tested by the Pearson product–moment correlation) indicates that the INR difference changes as INR values increase. We further tested the correlation between INR differences vs INR means, using Spearman nonparametric rank-order correlation.

The third parameter, or 95\% prediction interval, is the \(y\) (INR difference) range for a given \(x\) (INR average), where there is a 95\% probability that the next assay \(y\) value will occur based on the fit of the present experimental data. We used STATISTICA (data analysis software system), Ver. 6 (www.statsoft.com. StatSoft, Inc., 2004) throughout.

Results
We compared the results obtained by seven different commercial INR determination methods for 150 blood samples from patients with imminent or ongoing oral anticoagulation therapy. One sample gave very discrepant results with one reagent and was rejected. The final number of patients was therefore 149. All four ISI calibrators were used for the calibration of all seven methods. The INR values obtained with the different methods were very similar around the normal plasma value, or INR = 1 (Fig. 1). In contrast, marked differences were seen at higher INR values.

The results were dependent on the calibration method used (Fig. 1). Friedman ANOVA and Kendall concordance analysis demonstrated poor overall agreement between different methods when INR values obtained by individual calibrations (Fig. 1A) were assessed (\(P < 0.0001\); coefficient of concordance = 0.458). The means, medians, and standard deviations of the results are given in detail in Table 1. Pairwise comparison of individual INR results by the Wilcoxon matched-pairs test revealed significant differences between most pairs of assays (data
Fig. 1. Comparison of PT results obtained for 149 blood samples. 

(A), PT results obtained for 149 blood samples measured by seven different commercial PT methods with four calibrators. The methods (shown on the x axis) were as follows (respective calibrators are indicated in parentheses): 1, Innovin (Dade Behring); 2, PT-Fibrinogen Recombinant (Etaloquick); 3, PT-Fibrinogen HS Plus (Etaloquick); 4, Neoplastine CL Plus (Etaloquick); 5, Owren PT (Bioclin); 6, Nycotest PT (Axis-Shield); 7, SPA 50 (Etaloquick). (B–E), the same calibrator was used for all methods: (B), Bioclin; (C), Dade Behring; (D), Etaloquick; (E), Axis-Shield. To assist visual assessment, the data points representing different analyses of the same blood sample are connected by lines.
illustrated in Fig. 1A). Only 4 of 21 possible comparisons revealed nonsignificant differences between the two groups of INR results: the Innovin vs SPA 50 ($P = 0.146$); PT-Fibrinogen Recombinant vs the Owren PT ($P = 0.115$); PT-Fibrinogen HS Plus vs PT-Fibrinogen Recombinant ($P = 0.280$); and PT-Fibrinogen HS Plus vs the Owren PT ($0.340$).

Visual assessment of Fig. 1A revealed that the lack of a significant difference does not necessarily indicate similarity and that random and nonrandom variability may be present. The statistical significance of the differences in the other 17 comparisons persisted with even the most conservative statistical treatment using a Bonferroni-type correction, i.e., even after we multiplied the $P$ values by 17, they were still $<0.001$. This means that all INR methods tested here gave either inconsistently or systematically different results.

None of the four calibration methods, when applied to all INR methods, essentially harmonized the results (Fig. 1, B–E). This was also exemplified when we used Bioclin as a calibrator (Table 1). The maximum difference between the mean INR values obtained by two different PT methods (PT-Fibrinogen HS Plus and Neoplastine CL Plus) was 0.64 INR units.

We performed all 21 possible pairwise comparisons of the results obtained with the seven INR methods by plotting the results according to the Bland–Altman method. This strategy revealed three types of agreement problems: (a) The scatter of the difference in results obtained by the two methods to be compared increased according to the prolongation of PTs. In other words, the agreement between the two methods deteriorated toward the clinically important anticoagulation values. We observed a very large scatter ratio of the INR differences, 2.9 (Fig. 2J), with PT-Fibrinogen Recombinant vs the Owren PT. The absolute variation in differences is better indicated by the 95% prediction interval of the differences. In this regard, the most marked variation in agreement of results, i.e., a prediction interval of 2.9 INR units, was between PT-Fibrinogen Recombinant and Nycotest PT (Fig. 2J). Several comparisons revealed a substantial dependence of INR difference on the mean INR. This was indicated by the slopes of the regression lines in the Bland–Altman plots. The steepest slope, 0.54, was for Nycotest PT vs SPA 50 (Fig. 2U). All slopes differed statistically significantly from zero ($P < 0.001$, Pearson product–moment correlation analysis). Similarly, the nonparametric Spearman rank-order correlation test demonstrated a statistically significant correlation ($P < 0.001$) between the INR average vs INR mean.

**Discussion**

Harmonization of PT results and therapeutic ranges globally is an important goal. The introduction and recommendation of INR units were intended to serve this end because INR units are easy to handle by both patients and physicians and concordance of INR results worldwide would facilitate better care of individual patients. The major consideration was reliable and safe adoption of the results obtained by anticoagulation trials and of expert guidelines for treatment. The results of the present study, however, unfortunately demonstrate that the goal has not been achieved. Agreement among most of the INR methods tested here was particularly poor. Because of the wide global coverage of these reagents, our findings must be seen to refute the principle. In other words, the strategy of deriving equal INR results on the basis of the “WHO mother standard” (14, 15) does not work in its current form for manufacturer or for local calibrations.

The seven reagents selected all have low ISI values. The mathematical correction of INR results close to the ISI value of 1 is minimal. It is thus likely that the present study was conservative in revealing the agreement problems among PT reagents. The effect of ISI and the respective calibration, as a power function, increases with prolongation of PT times, so that even poorer agreement of INR results is likely with less sensitive reagents representing higher ISI values.

The results were in a very good agreement in patients with INR values close to 1, but increasingly poorer

| Table 1. Basic statistics of INR values obtained with different methods. |
|--------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Calibrator               | Innovin             | PT-Fibrinogen Recombinant | PT-Fibrinogen HS Plus | Neoplastine CL Plus | Owren PT            | Nycotest PT         | SPA 50              |
| Individual calibration‡  | Dade Behring        | Etaloquick            | Etaloquick            | Etaloquick          | Bioclin             | Axis-Shield         | Etaloquick           |
| Mean                     | 2.27                | 2.53                 | 2.45                 | 2.06                | 2.40                | 2.73                | 2.26                |
| SD                       | 0.72                | 1.00                 | 0.81                 | 0.66                | 0.70                | 1.07                | 0.62                |
| Median                   | 2.22                | 2.31                 | 2.42                 | 2.04                | 2.44                | 2.71                | 2.32                |
| Bioclin calibration‡     | Mean                | 2.36                 | 2.18                 | 2.61                | 1.97                | 2.40                | 2.44                |
|                          | SD                  | 0.87                 | 0.71                 | 0.93                | 0.54                | 0.70                | 0.72                |
|                          | Median              | 2.28                 | 2.04                 | 2.56                | 1.97                | 2.43                | 2.47                |
|                          |                     |                     |                     |                     |                     |                     |                     |
| a PT results, in INR units, obtained for 149 blood samples by use of seven different commercial PT determination methods and four calibrators.  
  b Statistics (mean, SD, median) for the data obtained by use of individual calibrators for each of the seven methods. The Dade Behring calibrator was used to calibrate Innovin, Etaloquick calibrator was used to calibrate PT-Fibrinogen Recombinant, and so forth. The original data are illustrated in Fig. 1A.  
  c Statistics (mean, SD, median) for the data in Fig. 1B; a universal calibrator (Bioclin) was used for calibration of all seven methods.  


agreement was evident with longer PT times. This applied to the scatter of results with most reagents. Furthermore, in most cases, the disagreement among results increased systematically toward longer PT times, and the recommended analytical bias \([\pm 0.20 \text{ INR units}]\) was clearly exceeded.

Can such marked divergence in results among PT methods be true? And what should be done? To answer
the first question, all available information must be taken into account. It must be noted that only a fraction of the blood components involved in the clotting reaction have been characterized. Different reagents are known to vary with respect to source of thromboplastins and other reagent components, and these have likewise not been characterized in detail. Hence, the theoretical basis of the ISI principle is itself not necessarily sound, as has been emphasized by several critics (17–19). There are, moreover, very few, if any, published investigations in which good agreement between different INR-based PT methods has been demonstrated. In contrast, the observations of Jackson et al. (28) as well as our own (21–23) have indicated poor agreement among PT methods. We emphasize, however, that the sampling in the current study was not necessarily performed in a chronic or stable phase of anticoagulation. Hence, it remains to be determined whether the remarkable bias among many INR methods applies similarly to samples taken from patients who have been on anticoagulant therapy for different periods of time.

How are we to improve the comparability of different INR methods? One, unfortunately impractical, answer would be to adopt one single reagent globally. A compromise strategy would be to use the same secondary standard in all laboratories. This approach, however, also is impractical. The next alternative would be for laboratories to choose the most appropriate PT determination method available. Our study demonstrates that, for the Quick PT, bias is greater and varies much more widely than for the Owren method. We did not investigate the biochemical mechanisms underlying this marked difference in the current study, but the theoretical reasons for the smaller variability of Owren PT times may well lie in the considerably smaller differences in the sample matrix (21, 29). In the Owren method, exogenous coagulation factor V and fibrinogen are added to the reaction mixture, and this eliminates the effect of the internal variation of these coagulation proteins. Furthermore, the proportion of the original plasma is only 5% in the Owren PT vs 33%, or a 6.6-fold higher proportion of plasma, in the Quick reaction. The substrate concentration is a well-known determinant of the velocity of the coagulation reaction (30). As shown here, the difference in INR results may of course be in equally poor agreement with longer Owren PT times. Nevertheless, the conceivably more appropriate reaction conditions and smaller INR value bias make the Owren PT principle an attractive alternative to achieve acceptable INR results for clinical practice.

In conclusion, we tested seven different INR methods and four calibrators, using 149 patient samples. Pairwise comparisons of the results demonstrated a clinically unacceptable divergence among most methods. The application of results of clinical anticoagulation trials and expert recommendations are thus by no means straightforward for clinical practice. Better approaches are needed. The Owren PT has some advantages over the Quick method, a theoretical consideration supported by the experimental data presented here.

This study was supported by the Tampere University Research Fund.

References


