Annals of the Rheumatic Diseases, 1987; **46**, 674–680

Standardisation of serological tests for rheumatoid factor measurement

F KLEIN[1] AND M B J A JANSSENS[2]

From the [1]Institute of Epidemiology, Medical Faculty, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands; and [2]T.N.O., Rheumatic Diseases Research Committee, Division of Health Research, Rijswijk, The Netherlands

SUMMARY Standardisation of quantitative data obtained by several types of rheumatoid factor test was achieved by the use of a reference serum preparation. Interlaboratory comparability improved for the latex fixation test, the Waaler-Rose test, the IgM RF test by an enzyme linked immunosorbent assay (ELISA), and for the antiperinuclear factor test. Use of a common method and latex preparation was not sufficient to improve comparability for the latex test. The comparability of IgM RF tests by immunofluoresence (IF) was not changed by reading against a common reference. It is concluded that expression in international units, as defined by the World Health Organisation (WHO), improves interlaboratory comparison of quantitative data in rheumatoid serology.

Key words: international units, reference preparations, antiperinuclear factor.

Rheumatoid serological tests, like most other biochemical determinations, are subject to a sometimes rather large variability between laboratories, dependent on methods and reagents. In one particular instance our own data showed a difference of six dilution steps between two laboratories in the same region for the same serum in the latex fixation test. Such discrepancies are not exceptional and could well result in different clinical judgments, or even cast doubt on the significance and reliability of quantitative data on rheumatoid factors (RF).[1] Efforts to reduce this variability by standardisation presuppose that quantitative data (e.g., titres) give more information to the clinician than a yes or no answer. This has indeed been demonstrated for RF.[2]

To improve comparability the WHO proposed a standard for RF determinations containing by definition 100 international units (IU) per ml.[3] Its effectiveness in the Waaler-Rose (WR) test system was better than in the latex fixation test (LFT). This standard seems to have gained little acceptance. Although Fulford *et al* reported a reduction of interlaboratory variance by the use of IU,[4] a later report stated that no improvement resulted in this way.[5]

In view of such conflicting data a Dutch working group studied the effectiveness of standardisation for various RF test methods. The possibility of standardising antiperinuclear factor (APF) determinations was investigated as well,[6] though at present there are no quantitative interpretations of this test. It appeared that in most cases standardisation was effective in reducing spread between laboratories. The results warranted an effort to introduce standardisation of RF tests in the Netherlands. As our experiences may be of interest to workers in the field of rheumatoid serology our methods and results are reported here.

Materials and methods

STANDARDISATION METHODS

In principle there are three alternatives for standardisation: (*a*) normalisation of procedures, (*b*) use of reference preparations, and (*c*) a central reference laboratory. We have opted for reference preparations, as normalised procedures present various practical problems of acceptance and central reference laboratories are not available. The effectiveness of reference preparations is based on

Accepted for publication 6 April 1987.
Correspondence to Dr F Klein, Institute of Epidemiology, Medical Faculty, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

the assumption that differences between laboratories are reflected in the same way by reference and test sera.

Results read against a reference serum can be expressed in twofold dilution steps ($\log_2$ reciprocal last positive dilution=tube number) or in relative units. Dilution steps are used for statistical purposes,[7] but units have been adopted for general laboratory practice. Dutch national units were based on the WHO international reference preparation 64/1, which contains 100 IU/ml by definition. The relation of units to dilution steps can be expressed by $\log_2 E_t - \log_2 E_s = \Delta t$ where E_t and E_s are the number of units in test and reference respectively and Δt is the number of twofold dilution steps between the end points of test and reference serum in the same run. Fig. 1 gives a schematic representation of this way of reading.

Each test system, including similar systems differing in IgG substrates only, was examined separately because antigenic differences between rabbit and human IgG as well as differences in sensitivity between detection methods may produce essentially incomparable results. For the same reason units in reference preparations have been assigned to each test, adjusting the WHO reference to 100 IU/ml in every instance.

The WR test, the LFT, the detection of IgM RF by immunofluorescence[8] or by ELISA (various methods) and the APF test were analysed separately. The IgG RF tests were not examined because clinical studies cast serious doubt on the relevance of this test compared with the IgM RF test.[9]

The effectiveness of standardisation by a reference serum was tested by comparing interlaboratory spread in groups of laboratories before and after standardisation. In accordance with our chosen principle, no attempt was made for all participants in comparative studies to use the same test method. Only a standardised LFT procedure[10] was used by every member of the working group, besides his own test system, in order to study the effect of a standardised method and reagent in one instance. Five or six laboratories, being members of the Dutch working group, participated in all comparative experiments. In one experiment 33 laboratories took part in order to study the effect of a proposed reference preparation on interlaboratory comparability on a larger scale.

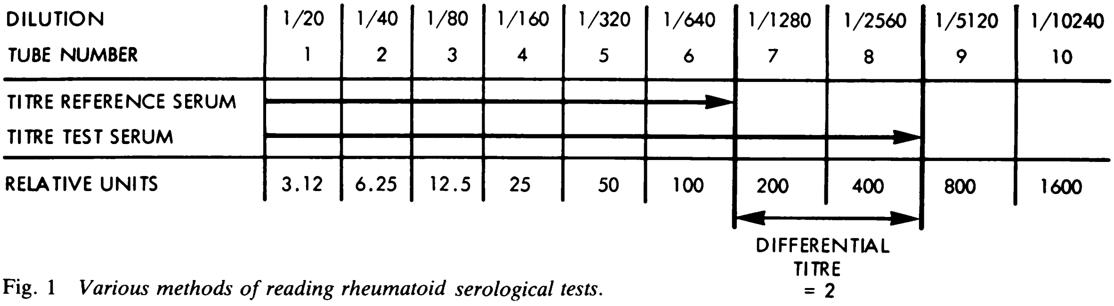
In comparative studies participants were asked to analyse a number of the same test sera together with a common reference in their own test system, sometimes on more than one day, and to express their results as titres. Each test serum and reference preparation had to be titrated until a negative reaction was obtained. The proposed common reference was first calibrated by reading it against the WHO preparation 64/1 as a primary reference. The reference value in IU was calculated according to the formula $E(\text{ref})=2^{\Delta t}\times 100$ where E (ref) is the number of units in the proposed Dutch reference and Δt the median difference in positive twofold dilution steps between the Dutch and the WHO reference. In all experiments the participants knew which samples were the reference sera.

STATISTICAL ANALYSES

The results of the tests before standardisation were expressed as $t=\log_2$ reciprocal dilution and after standardisation as $\Delta t=\log_2$ reciprocal dilution of reference$-\log_2$ reciprocal dilution of test serum.

For statistical analyses differing dilution series had to be converted to one standard series, for which 1/2, 1/4, 1/8 . . . was chosen, with $\log_2$ reciprocal dilution=1, 2, 3. . . . For other dilution series $\log_2$ reciprocal dilution=$\log_{10}$ titre/$\log_{10}$ 2. Thus one may obtain fractional numbers, e.g., 2·3, for dilution 1/5. As all participants used the series 1/20, 1/40, 1/80 for the LFT this series was transformed to 1, 2, 3. . . .

These figures were analysed by the classical analysis of variance method with the transformed titre as dependent variable and laboratories, test sera, and days as independent variables. This



DILUTION	1/20	1/40	1/80	1/160	1/320	1/640	1/1280	1/2560	1/5120	1/10240
TUBE NUMBER	1	2	3	4	5	6	7	8	9	10
TITRE REFERENCE SERUM										
TITRE TEST SERUM										
RELATIVE UNITS	3.12	6.25	12.5	25	50	100	200	400	800	1600

Fig. 1 *Various methods of reading rheumatoid serological tests.*

analysis produces the test statistic F=s^2 between laboratories/s^2 residual. The BMD.P2V computer program[11] was used for calculations. When tests were carried out on more than one day the variance between days was not extracted from the residual variance in the denominator of the F test. In this way the F test was able to demonstrate to what extent the variance between laboratories exceeded the variance between days. Ideally the F value should become non-significant after standardisation. In practice this is rarely the case. If standardisation reduced the spread between laboratories, the F value should become smaller after standardisation than before. The significance of this decrease can be tested by calculating the test statistic F′=(s^2 between laboratories−s^2 labs sera) before standardisation/(s^2 between laboratories−s^2 labs sera) after standardisation, where s^2 labs sera is the interaction variance between laboratories and test sera (van Strik, personal communication). This is an approximation of Pitman and Morgan's test for correlated variances. The variances between laboratories and the interaction variances in this formula were obtained from the BMD.P2V program.

The analysis of variance is based on certain assumptions about normality. These were not explicitly tested, but as a general check a few analyses were repeated using a non-parametric method.[12]

SERA

Test sera were obtained from individual patients with rheumatoid arthritis (RA). In several comparative studies one or more test sera were supplied in duplicate to the participants, without their knowledge. This served as a control to eliminate gross cases of poor reproducibility, which however hardly occurred. Such serum samples were treated as independent variables in the analyses of variance.

NETHERLANDS REFERENCE SERUM PREPARATION (NRSP)

Preparation and stability

Plasma was obtained by plasmapheresis of about 20 patients with rheumatoid arthritis and converted to serum by recalcification. The serum samples were examined for RF by various techniques, and for APF. A pool was made in such a way that it would contain a number of IU not too far removed from the WHO standard. The pool was defatted with Freon, transferred to ampoules, and freeze dried. Sodium azide was added as a preservative to a final concentration of 0·033%. The stability of this preparation was tested by keeping 450 ampoules in groups of 15 at −196°C, −20°C, 4°C, 20°C, and 37°C for periods of up to one year, testing them at regular intervals. Even after one year at 37°C the titres dropped less than one dilution step. Keeping a solution of the preparation frozen at −20°C resulted in a significant titre decrease after three months.

Variability and calibration of the NRSP

Three batches of the NRSP were analysed in duplicate by six laboratories on three different days, each day together with the WHO primary reference 64/1 in duplicate. By analysis of variance it was established that for all tests the SD between batches and between days was nearly always less than one twofold dilution step (see Table 1). This means that there was no appreciable difference between batches and between days, and therefore the NRSP is homogeneous and gives reproducible results. The F values were not significant, or in other words the calculated interday and interbatch variances were not significantly greater than the random (residual) variances. The only exception (IgM RF between days) remains unexplained.

The NRSP was calibrated by comparison with the WHO reference preparation 64/1 (containing 100 IU/ml) for each test system and substrate. The values in IU of the NRSP for various systems are summarised in Table 2.

TESTS FOR INTERLABORATORY COMPARABILITY

Preliminary results with the proposed NRSP

In this experiment the effect of using a reference on interlaboratory comparability was tested on a selected group of participants (the Dutch working group mentioned before). Five test sera were analysed once on three different days by four to six laboratories (depending on the test) together with the NRSP in sixfold replicates. All participants used the same twofold dilution series, but their own methods, except for the latex-Norde test[10] and the immunofluorescent tests for IgM RF,[8] which were performed by all participants in the same way. A mixed data set for the LFT was composed of participants who only used the latex-Norde test and those who had used other latex preparations and methods. One test serum was found negative by all laboratories and was therefore omitted from the calculations, as it would reflect only the variation of the reference.

The results of this experiment were examined by analysis of variance and by Tukey's non-parametric method. The total spread rarely exceeded two dilution steps after standardisation. In the classical analysis F became smaller but remained significant at the 0·005 level or lower after standardisation for all tests (Table 3). The F′ test statistic (see 'Methods')

Table 1 *Standard deviations of transformed titres and F values between days and between batches for the Netherlands reference serum preparation*

Test	*Between days*			*Between batches*		
	SD	*F*	*p*	*SD*	*F*	*p*
Latex (various methods)	0·54	1·82	0·19	0·27	0·45	0·64
WR	0·52	2·00	0·17	0·26	0·50	0·62
IgM RF (IF)	1·15	5·43	0·02	0·76	2·38	0·13
APF	0·33	1·26	0·32	0·44	2·21	0·15

Table 2 *Values in IU/ml of the Netherlands reference serum preparation in several detection systems for RF and APF*

Test system	*NRSP (IU/ml)*
Latex	200 (all tests systems)
WR	400 (all test systems)
IgM RF	200 (immunofluorescence, rabbit gammaglobulin)
IgM RF	200 (ELISA, rabbit gammaglobulin)
IgM RF	200 (ELISA, human gammaglobulin)
APF	200

has been used to judge the significance of the decrease of F. In subsequent analyses only F′ has been given.

It can be seen that F′ was significant in the case of the mixed latex test data and of the APF but not for the IgM RF and the WR test. It should be noted that IgM RF was determined by comparable methods by all laboratories but that use of the same method and latex reagent (latex-Norde test[10]) did not sufficiently diminish the differences between the latex tests. This can be inferred from the calculated value of $F=66{\cdot}7$ ($p<0{\cdot}0001$) before standardisation. The results suggest that reading the RF and APF tests against a common reference will improve the comparability between laboratories. A non-parametric analysis of the figures confirmed this conclusion (data not shown).

This experiment, although encouraging, is not yet sufficient to conclude that our method of standardisation is effective. The participants of the working group might have been too familiar with each others' methods as in the WR and IgM RF tests and too well trained in comparative studies of this kind. A more unbiased experiment will be described in the next section.

Test with a large group of participants

To test the effectiveness of standardisation on a larger and more representative group another experiment was carried out. The participants were 33 laboratories engaged in routine rheumatoid serology (regional and hospital laboratories, as well as the members of the working group). Only 11 of these laboratories carried out the APF test and seven the immunofluorescent test for IgM RF.

Each laboratory received 10 sera which were actually five duplicates without the recipients being aware of it. These sera had to be analysed together with the NRSP in fourfold replicates on one day only. The participants used their own methods and dilution series. The only common obligation was to titrate all sera until a completely negative result was obtained.

A problem arose when it appeared that the BMD.P2V computer program was unable to handle blocks much larger than 10×10. Therefore the WR test and the latex test had to be analysed in three almost equal parts, each containing a comparable number of high and low titre values. Analysis of variance was carried out as usual on each subset with F′ as test statistic. The results are summarised in Table 4, and Fig. 2 gives as an example the scatter diagrams obtained for one particular serum in the latex test before and after standardisation. The difference between two laboratories in the same region (X and Y) decreased from six to one dilution step after standardisation.

Although most of the participants had no experience with standardisation and some of them little even with titration, the effect of reading against a

Table 3 *Analysis of variance of comparative tests for RF and APF*

Test	*F*	*p*	*F*	*p*	*F′*	*p*
	(before stand)		*(after stand)*			
LFT-Norde	66·7	<0·0001	11·5	<0·0001	13·90	0·01<p<0·05
All LFT	733·7	<0·0001	13·5	<0·0001	80·32	<0·01
WR	11·8	<0·0001	4·7	0·0036	3·72	0·10<p<0·25
IgM RF	7·7	0·0005	5·2	0·0050	1·86	>0·25
APF	89·8	<0·0001	9·5	0·0001	17·44	0·01<p<0·05

Table 4 *Analysis of variance of comparative tests for RF and APF*

Test	*n*	*F'*	*p*
Latex			
1	9	41·75	<0·01
2	9	25·27	<0·01
3	9	9·27	<0·01
WR			
1	10	5·47	<0·01
2	8	12·67	<0·01
3	11	7·21	<0·01
IgM RF	7	1·04	>0·25
APF	11	7·39	<0·01

n=number of participants.

Table 5 *Standardisation of the ELISA test for RF*

	F'	*p*
Rabbit gammaglobulin	199·00	<0·01
Human gammaglobulin	5·59	0·01<p<0·05

common reference was striking, also for the WR test, but again with the exception of the IgM RF test. The SD after standardisation never exceeded 1·2 dilution steps. The split-up data all behaved in the same way, thereby proving the correctness of this procedure. Further confirmation was obtained from analysis by a non-parametric method (data not shown). This experiment showed that standardisation was also effective outside specialised laboratories.

Standardisation of ELISA test systems

These test systems were used with rabbit and with human gammaglobulin as substrates, and with antihuman IgM in the detection system. Both substrates were analysed separately. The content of the NRSP in international units was assessed for each substrate separately, but was the same in both cases (see Table 1).

Six laboratories took part in the investigation, each with its own method. They received four test sera and one reference (NRSP), to be analysed in one run only. The results in the ELISA method are usually regarded as a continuous variable and read on a calibration line of a standard. When read in this way the results are already standardised, though possibly not on the same standard. As comparison of results would seem senseless in this way the effect of standardisation was calculated from titres which every participant had to determine for this occasion. Comparison of titre end points and differential titres (Δt) allowed an estimation of interlaboratory spread before and after standardisation, as in previous experiments.

The results were analysed as usual and are summarised in Table 5. The tests with rabbit gammaglobulin seemed to be more effectively standardised than those with human gammaglobulin.

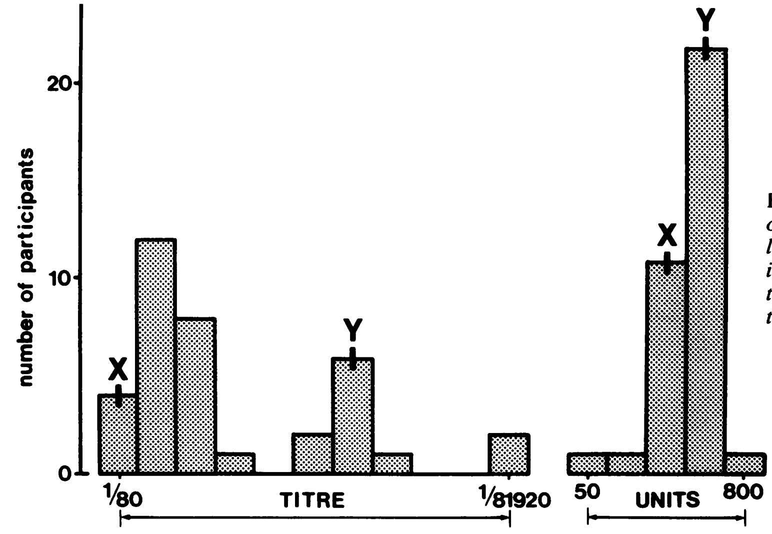


Fig. 2 *Results of the latex test in one test serum by several laboratories, expressed in titres and in international units. X and Y are two different laboratories in the same region.*

Discussion

The need for standardisation is often insufficiently understood in the field of rheumatoid serology. In fact all measurement should be followed by standardisation if the results are going to be of any general use. This is a commonplace for measurements of length, for example, but not yet for serological tests. Critical comparison of kits and reagents for assays also requires previous standardisation of results as soon as more than one laboratory is involved.

The present investigation shows that standardisation of WR, LFT, IgM RF tests by ELISA, and also APF tests improves interlaboratory comparability. This is of obvious importance for exchange of scientific information and also for the clinician who may be confronted with discrepant data for the same patient, obtained from different laboratories. The case illustrated in Fig. 2 shows that the doubts of Lea and Ward[1] concerning seropositivity may sometimes be justified. Expression in international units diminishes such problems.

Although perfect comparability was not obtained in our study, a smaller spread than we found can hardly be expected in view of the large number of participants, methods, and reagents involved. Deviating results can only be traced by continuous external quality control, a necessary complement to standardisation.

The WHO study showed disappointing results for the LFT,[3] which in our hands, however, was effectively standardised. Standardisation by the use of one 'suggested method' by every participant was tried by Anderson *et al* for the sheep cell agglutination test and proved to be ineffective.[3] In the present investigation similar negative results were obtained with a standardised LFT. Therefore there seems to be little hope for a standardised latex procedure as proposed by Singer.[13] Standardised methods (and reagents!) would also pose problems of acceptance.

Although Fulford *et al* obtained a reduction of variance between laboratories by using a common reference,[4] Rippey and Biesecker did not find this in a later report.[5] From their own figures, however, a favourable effect of reading in relative units still appears when it is realised that results in IU form a geometric progression and should be compared in that form with titres, which Rippey and Biesecker did not. Our own work provides solid support for standardisation by reference sera, based on a larger number of test sera and on statistical evidence for the effectiveness of this method of improving comparability of data.

The lack of acceptance of IU until now may be due to a failure to realise the possible range of scatter without standardisation. Scatter diagrams like that of Fig. 2 show that giving a titre is quite an arbitrary way of expressing the results of RF assays. In the field of immunoglobulin measurement IU have also been accepted reluctantly, if at all. It is becoming clear now, however, that absolute immunoglobulin measurement (in mass units) is possible with acceptable precision.[14] In RF serology this is not possible because of very different substrate affinities associated with the same mass of IgM RF. It is therefore to be hoped that the use of IU will gain ground in this field.

In calibrating the Dutch reference serum preparation we have assumed that WHO reference 64/1 still has the same potency; Hay and Nineham have raised doubts about this.[15] Such doubts may interfere with general acceptance. This, however, should not keep clinical and other laboratories from paying more than lip service to accepted principles of biological standardisation.

This study was carried out by the Dutch working group on standardisation of rheumatoid serology. The following have participated in this study: J H de Bruyn, H A L Clasener, E Desser, T E W Feltkamp, H G M Geertzen, I Jankowski, C Kamphorst, J Katchaki, J A M Kerckhaert, G J M Lafeber, P Limburg, L K J van Romunde, J A M Snyder, N Verwey-Burke, M L Westedt. Special thanks are due to Professor R van Strik (Biostatistics Dept, Erasmus University, Rotterdam) for his advice on statistical evaluation. Financial support was received from the Foundation Reference Laboratory for Rheumatoid Serology (RELARES).

References

1 Lea D J, Ward D J. The case against 'seronegative RA': a laboratory viewpoint. *Br J Rheumatol* 1984; **23:** 236.

2 Cats A, Klein F. Quantitative aspects of the latex fixation and Waaler-Rose test. *Ann Rheum Dis* 1970; **29:** 663–72.

3 Anderson S G, Bentzon M W, Houba V, Krag P. International reference preparation of rheumatoid arthritis serum. *Bull WHO* 1970; **42:** 311–8.

4 Fulford K M, Taylor R N, Przybyszewski V A. Reference preparations to standardise results of serological tests for rheumatoid factor. *J Clin Microbiol* 1978; **7:** 434–41.

5 Rippey J H, Biesecker J L. Results of tests for rheumatoid factor on CAP survey specimens. *Am J Clin Pathol* 1983; **80** (suppl): 599–602.

6 Sondag-Tschroots I R J M, Aay C, Smit J W, Feltkamp T E W. The antiperinuclear factor. I. The diagnostic significance of the antiperinuclear factor for rheumatoid arthritis. *Ann Rheum Dis* 1979; **38:** 248–51.

7 Singer J M, Edberg S C, Selinger M, Amram M. Quality control of the latex-fixation test. *Am J Clin Pathol* 1979; **72:** 591–6.

8 Estes D, Atra E, Peltier A. An immunofluorescent method for the detection of antigammaglobulin antibodies. *Arthritis Rheum* 1973; **16:** 59–65.

9 Westedt M L, Herbrink P, Molenaar J L, *et al.* Rheumatoid factors in rheumatoid arthritis and vasculitis. *Rheumatol Int* 1985; **5:** 209–14.

10 Klein F, Bronsveld W, Norde W, van Romunde L K J, Singer J M. A modified latex fixation test for the detection of rheumatoid factors. *J Clin Pathol* 1979; **32:** 90–2.

11 Jennrich R, Sampson P, Frame J. Amalysis of variance and covariance including repeated measures. In Dixon W J, ed. *BMD statistical software*. California: University of California Press, 1981: 359–87.

12 Tukey J W. *Exploratory data analysis*. Reading, Mass: Addison-Wesley, 1977: 363–400.

13 Singer J M. Standardization of the latex fixtion test for rheumatoid arthritis serology. *Bull Rheum Dis* 1974; **24:** 762–9.

14 De Bruyn A M, Klein F, Neumann H, Sandkuyl L A, Vermeeren R, Le Blansch G. The absolute quantification of human IgM and IgG: standardization and normal values. *J Immunol Methods* 1982; **48:** 339–48.

15 Hay F C, Nineham L J. Standardization of assays for rheumatoid factors and antiglobulins. In: Dumonde D C, Stewart M W, eds. *Laboratory tests in rheumatoid diseases*. Lancaster: MTP Press, 1979: 101–5.