Scand J Work Environ Health 1992;18:246—51

Agreement on medical fitness for a job

by Wim LAM de Kort, MD,[1] Hein W Post Uiterweer, MD,[2] Frank JH van Dijk, MD[3]

DE KORT WLAM, POST UITERWEER HW, VAN DIJK FJH. Agreement on medical fitness for a job. *Scand J Work Environ Health* 1992;18:246—51. Five experienced occupational physicians independently reviewed the uniformly structured, concise records of 180 applicants who had applied for a job in one of three categories. All had undergone a preemployment medical examination by the Governmental Occupational Health and Safety Service. Agreement was assessed by calculating the percentage of disagreement and Cohen's kappa. Agreement between the five panel physicians and between the panel physicians and the Service appeared to be poor, with overall percentages of disagreement of 31 and 37%, respectively, and kappa values of 0.38 and 0.37, respectively. On the average 31% of the applicants judged as unfit by one physician had been assessed as fit by the others, whereas agreement was only marginally better when detailed medical criteria for fitness were available. Lack of consensus on the medical fitness of an applicant, as evidenced by this study, suggests that the validity of such a judgment may be questionable even when detailed fitness criteria are available.

Key terms: health policy, interobserver variability, kappa statistic, occupational health care, personnel selection, preemployment medical examination.

Preemployment medical examinations are carried out in many industrialized countries. The aim of such examinations is to judge the medical fitness of an applicant for a certain job (1—6). In an attempt to define this aim more precisely, however, a descriptive inventory among physicians in The Netherlands revealed that, at least in that country, large differences of opinion exist (7). In contrast, the content of preemployment medical examinations and the references used appear very similar in different countries (1—6). The pros and cons of having preemployment medical examinations are regularly debated (8—16), while their actual effectiveness is seldom discussed (17—19), and no literature is available on the validity of decisions arising from preemployment medical examinations.

A previous study analyzed the results of preemployment medical examinations in a well-defined applicant population seen by the Governmental Occupational Health and Safety Service (GOS) over a six-year period (1983—1988) (20). It was found that about 20% of the applicants had a medical condition of a major or minor ailment or disorder. About 0.6% of the applicants were judged by the GOS to be unfit for specific jobs. The rejection percentages varied from around 0.3% for administrative personnel up to 3.5% for security personnel (eg, prison officers). The rejection percentages for medical diagnostic categories showed a markedly smaller variation than the rejection percentages for job categories, a finding suggesting that job category is a stronger determinant for rejection than medical diagnostic category.

Although for many job categories rejection percentages appear to be low (7, 17, 20), the correct identification of medically unsuitable applicants is still important because accepting a medically unfit applicant ("false negative") may result in unfortunate consequences for the applicant and the employer. Equally unfortunate is the inappropriate rejection of a suitable applicant on medical grounds ("false positive").

Whether an applicant is validly judged unfit is difficult to assess, since randomized, longitudinal study designs are not appropriate due to the great methodological and ethical problems. Good agreement in the interpretation of medical test results, however, is a prerequisite for a valid judgment. We have, therefore, studied the degree of agreement between experienced occupational physicians in considering the medical fitness of applicants.

Subjects and methods

Study population

The study population consisted of 180 applicants, drawn from a population of 101 754 who had applied for various jobs in governmental service in the years 1983—1988. They had been seen for a preemployment medical examination by the GOS. (See table 1.) Only the GOS records of those applicants who had applied for jobs in one of three categories, namely, administration, cleaning and catering or prison security, were used.

[1] Medical Biological Laboratory TNO, Rijswijk, The Netherlands.

[2] Governmental Occupational Health & Safety Service, The Hague, The Netherlands.

[3] Coronel Laboratory, University of Amsterdam, Amsterdam, The Netherlands.

Reprint requests to: Dr WLAM de Kort, Medical Biological Laboratory TNO, PO Box 45, 2280 AA Rijswijk, The Netherlands.

For each job category 30 records were randomly selected from the group of applicants judged to be unfit or temporarily unfit (ie, were considered doubtful cases), yielding 90 records of applicants, henceforth called cases. For each case the medical diagnosis that had led to rejection was noted. For each case record, a reference record was selected of an applicant who had been judged fit for the same type of job. For both the reference applicant and the case applicant a corresponding medical diagnosis should have been made by the GOS.

Procedure

Two of the authors (WdK, HPU) prepared uniformly structured, concise reports of all 180 records containing a job specification, questionnaire results, medical examination data and test results, and other information, if present. All of the information was retrieved from the medical records. Each record contained a job specification form, a comprehensive medical questionnaire (filled out by the applicant), and a medical examination registration form (filled out by the physician). Many (but not all) of the records contained additional information, such as correspondence with general practitioners or medical specialists, psychological reports, specialized test results, or laboratory data. In addition to the concise reports, a comprehensive job specification and the detailed medical fitness criteria used for prison officers by the GOS were made available to the panel physicians.

Five physicians, who were officially registered as occupational physicians for at least nine years and for whom carrying out preemployment medical examinations was a routine task, were invited to take part in the study. Each one independently reviewed the concise reports and indicated if, in their judgment, the applicant was fit or unfit for the job for which he or she had applied. Alternatively, given the available data, they might also decide that no conclusion could be reached. They received the reports in a random sequence, were unaware of the judgment of the GOS, and had three months to complete the task.

Statistical analysis

Two tests for pairwise interobserver agreement were used: (i) the percentage of judgments for which disagreement existed, a crude agreement measure confounded by coincidental agreement, and (ii) Cohen's kappa or kappa, a statistic in which a correction for coincidental agreement is made (21, 22). Kappa can have values from −1 to +1, where a value of 0 indicates pure coincidental agreement, values of 0—0.39 indicate poor agreement, values of 0.40—0.59 indicate moderate agreement, values of 0.60—1 indicate good agreement, and a value of +1 indicates perfect agreement. Negative values correspond with opposite opinions.

In accordance with Feinstein's suggestion (23), the average percentage of disagreement and the kappa values of all possible (ie, 10) pairwise comparisons between panel physicians) were used as the group measure for interobserver agreement among them. The average percentage of disagreement and the kappa values of all possible (ie, five) pairwise comparisons between the panel physicians on one hand and the GOS on the other were used as the group measure for interobserver agreement between the panel physicians and the GOS. The BMDP software package was used for all of the statistical analyses (24).

Results

Judgments of the Governmental Occupational Health and Safety Service and panel judgments

The 90 referents had, by definition, been judged as fit by the GOS (50% of the study population). Of the 90 cases, 22 had been judged as temporarily unfit (doubtful cases 12%). After a reexamination, 17 of these 22 were deemed medically fit. Five doubtful cases were lost to the follow-up; they had not been appointed to the job and had not been reexamined by the GOS. Of the remaining 68 cases (38%) who had been judged

Table 1. Summary of the source population data. The study population was drawn from the job categories of lower administrative personnel, prison officers, and cleaning and catering personnel.

Job category	N	Age mean (years)	Men (N)	Women (N)	With a medical condition (%)	Doubtful cases (%)	Unfit (%)
Administrative personnel	22 762	25.3	39	61	22.3	0.14	0.16
Prison officers	2 608	29.1	76	24	27.8	0.84	2.80
Cleaning and catering personnel	2 254	29.1	32	68	24.0	0.35	1.06
All job categories (including the above)	101 754	27.5	58	42	20.6	0.17	0.43

as unfit by the GOS, 36 had appealed against this decision. After reexamination by a Committee of Appeal, 23 were declared fit.

Therefore, the final result of the GOS procedure, including the procedure of appeal, revealed that, of the 180 applicants in the study population, 130 (90+17+23) applicants (72%) were assessed as fit for the job by the GOS and 45 (32+13) applicants (25%) as unfit. In the case of five applicants (3%) the GOS evaluation of a "doubtful case" had not been reassessed and they remained doubtful cases.

The panel physicians, on the average, judged 67 (range 57—74)% of the 180 applicants as fit for the job and 20 (range 14—27)% as unfit. For 13 (range 8—17)% no conclusive judgment was reached (ie, they were judged doubtful cases). (See figure 1.)

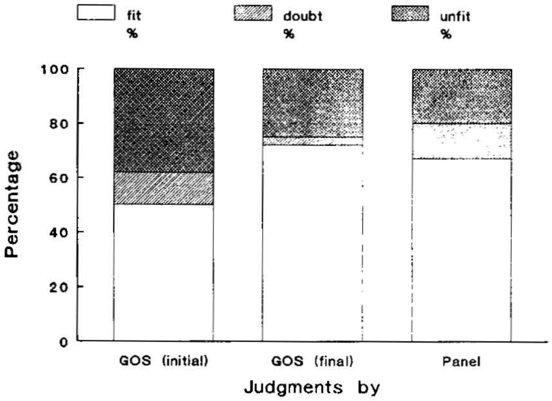

Figure 1. Results of preemployment medical examinations in the study population. The (final) results of the Governmental Occupational Health and Safety Service (GOS) refer to the results after reexaminations of the applicants who were judged to be temporarily unfit and the applicants who were judged to be unfit but who appealed against this initial GOS decision.

Table 2. Agreement matrix of the comparisons between the panel physicians and the Governmental Occupational Health and Safety Service (GOS). Each matrix cell percentage is the calculated average of the corresponding matrix cell percentages of the five individual panel physician and GOS matrices ($N_{total} = 180$).[a]

GOS decision	Panel judgment			
	Fit (%)	Doubt (%)	Unfit (%)	Total (%)
Fit	43	4.0	3.5	50
Doubt	8.3	3.8	0.1	12
Unfit	16	5.4	17	38
Total	67	13	20	100

[a] Kappa = 0.35, percentage of disagreement = 37.

Table 3. Matrix for the agreement among the panel physicians. Each matrix cell percentage shown is the calculated average of all corresponding matrix cell percentages of the 10 individual physician A and physician B matrices ($N_{total} = 180$).[a]

Judgment of physician B	Judgment of physician A			
	Fit (%)	Doubt (%)	Unfit (%)	Total (%)
Fit	54	6.7	6.3	67
Doubt	6.7	3.8	2.6	13
Unfit	6.3	2.6	11	20
Total	67	13	20	100

[a] Kappa = 0.38, percentage of disagreement = 31.

Agreement between the panel physicians and the Governmental Occupational Health and Safety Service and between the panel physicians

In this section only the initial GOS results have been taken into consideration (ie, before any reexamination). Tables 2 and 3 show the agreement matrices of the comparisons between the panel physicians and the GOS and also the comparisons between the panel physicians' judgments. With regard to those applicants who had been judged doubtful cases, no straightforward interpretation of the differences in judgments was possible. Of those 68 applicants who had been considered unfit by the GOS, on the average, 42% had been judged as fit by the panel physicians. This calculated average can be considered the percentage of "false positives": 16/38 = 42% (range 29—63%, depending on the job category). (See table 4.) Of the 90 applicants assessed as fit by the GOS, on the average, 7.0% had been judged as unfit by the panel physicians and can be considered missed cases or "false negatives": 3.5/50 = 7.0% (range 0.7—16% depending on the job category).

Among the panel physicians the percentage averaged 31 (range 25—44)% for false positives and 9.4 (range 4.6—16)% for false negatives. (See table 5.) These subpopulations of false positives and false negatives were not identical in composition per individual physician. For example, in only nine cases did all panel physicians agree that an applicant should be judged as unfit, while in 70 cases they unanimously agreed that an applicant should be deemed fit.

These rather substantial disagreements were reflected in the percentages of disagreement and the kappa values. For all 180 applicants taken together, between the panel physicians and the GOS, the percentage of disagreement was 37 (SD 4)%, and the kappa was 0.35 (SD 0.07). For the panel physicians, the percentage of disagreement was 31 (SD 4)% and the kappa was 0.38 (SD 0.06).

Differences in agreement between job category subgroups are of special interest, since, for prison officer applicants, detailed fitness criteria were made available. However, of the applicants for prison work who had been judged as unfit by the GOS, 29% were considered fit by the panel physicians. Conversely, of the applicants for prison work who had been judged fit by the GOS, 16% were deemed unfit by the panel physicians. The figures for comparisons between the panel physicians were 25 and 16%, respectively. See table 5.

The percentages of disagreement and the kappa values were separately calculated for each job category; see table 6 (the corresponding agreement matrices are not shown). With regard to the group of prison officer applicants agreement indeed appeared to be better than within the two other job categories. However, the agreement among panel physicians on the fitness

Table 4. Percentages of applicants with a panel physicians' judgment equal to or different from the judgment of the Governmental Occupational Health and Safety Service (GOS). For example, of the administrative personnel applicants judged fit by the GOS, the panel physicians judged 0.7% unfit and they had doubt about another 7.3%. The totals have been directly calculated from table 2.

Job category	GOS judgment								
	Fit (N = 90)			Unfit (N = 68)			Doubt (N = 22)		
	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)
Administrative personnel	92	0.7	7.3	47	33	20	73	0.0	27
Prison officers	76	16	8.0	29	60	12	60	6.7	33
Cleaning and catering personnel	87	4.7	8.7	63	25	13	65	0.0	35
Total	85	7.0	8.0	42	44	14	69	0.9	31

[a] Panel physicians' judgment.

Table 5. Percentages of applicants with judgments equal to or different among panel physicians. For example, of the administrative personnel applicants judged fit by panel physician A, physician B judged 4.6% unfit and had doubt about another 11%. The totals have been directly calculated from table 3.

Job category	Physician A's judgment								
	Fit (N = 121)			Unfit (N = 36)			Doubt (N = 23)		
	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)	Fit[a] (%)	Unfit[a] (%)	Doubt[a] (%)
Administrative personnel	85	4.6	11	32	47	22	54	15	31
Prison officers	75	16	8.6	25	64	12	42	37	22
Cleaning and catering personnel	81	9.1	10	44	46	10	54	11	35
Total	81	9.4	10	31	56	13	51	20	30

[a] Physician B's judgment.

Table 6. Agreement among the panel physicians and agreement between the panel physicians and the GOS. (A = all 180 applicants, B = stratified for job category, C = stratified for cases and referents, D = stratified for two GOS-diagnosed disorders, GOS = Governmental Occupational Health and Safety Service)

Stratum	Panel				Panel versus GOS			
	Kappa	SD[a]	Percentage of disagreement	SD[b]	Kappa	SD[a]	Percentage of disagreement	SD[b]
A								
All applicants (N = 180)	0.38	0.06	31	4	0.35	0.07	37	4
B								
Administrative personnel (N = 60)	0.34	0.09	27	6	0.31	0.08	39	4
Prison officers (N = 60)	0.40	0.09	35	6	0.41	0.12	33	7
Cleaning personnel (N = 60)	0.32	0.07	31	7	0.30	0.12	40	5
C								
Referents (N = 90)	0.28	0.12	19	4	·	·	14	5
Cases (N = 90)	0.32	0.07	43	5	·	·	59	9
D								
Disorder of the musculoskeletal system (N = 50)	0.38	0.11	34	7	0.34	0.14	36	6
Psychological/psychiatric disorder (N = 24)	0.24	0.09	42	1	0.33	0.22	39	15

[a] SD of Kappa.
[b] SD of percentage of disagreement.

of prison officer applicants was statistically significantly better only in comparison with cleaning and catering applicants. All in all, even the agreement on the prison officer applicants remained at a moderately low level.

To elaborate the poor agreement, we divided the study group into subgroups by stratifying for cases and referents, as defined in this study, and by stratifying for diagnostic categories as assessed by the GOS. (See table 6.)

When the agreement on cases was compared with that on referents, the percentage of disagreement of the referents appeared to be much lower than the percentage of disagreement of the cases, while (among the panel physicians) the kappa values did not show significant differences. Obviously, due to the unequal number of possible decision categories, no kappa value could be calculated when the agreement of cases and referents was assessed between the panel physicians and the GOS.

The diagnostic categories which contained enough applicants to justify statistical analysis were "disorders of the musculoskeletal system" and "psychological/ psychiatric disorders." The degree of agreement on the fitness of applicants did not differ significantly between these diagnostic categories (ie, the kappa showed low values for both groups).

Discussion

The agreement on which applicants should be judged medically fit or unfit was rather poor among the panel physicians, as well as between the panel physicians and the GOS. The highest and lowest rejection percentages differed by a factor of two among the five panel physicians. However, the situation in which the panel physicians had to judge applicants was artificial in several ways. The judgments had to be made on summarized record data, the panel physicians had no personal contact with the applicants, and the study population was not a random sample of the source population.

The data from the GOS examinations, as well as other information, for example, correspondence with general practitioners or medical specialists, had been summarized. Therefore, a "total view" of the applicants' records had not been available. Moreover, no personal interactive contact had been possible. Part of the disagreement in judgment between the panel physicians and the GOS might be attributable to these factors. However, the panel physicians all had exactly the same data at their disposal. Therefore, disagreement among them cannot be attributed to differences in information. Moreover, there is no evidence that agreement in real life situations would have been much better, considering the results of the examinations and reexaminations of the GOS. Twenty-three applicants who at first had been found to be unfit from a group of 36 were, on reexamination by the Committee of Appeal, in fact declared medically fit for the job. This Committee of Appeal again had personal contact with the applicant involved.

In practice, for about 20% of the applicants a diagnosis of a major or minor ailment or disorder had been noted by an examining GOS physician, while in the three job categories only 0.3—3.8% of the applicants had been judged (temporarily) unfit by the GOS. (See table 1.) In the study population a medical condition had been noted in all of the applicants by an examining GOS physician and the percentage of applicants judged (temporarily) unfit by the GOS was much larger, namely, 50%. Because of the overrepresentation of applicants with a medical condition and of applicants judged to be unfit, the panel physicians may conceivably have been inclined to judge a lower percentage of applicants as unfit, due to a "regression to the mean" phenomenon. Again, however, this phenomenon does not explain the poor agreement among the panel physicians.

In a comparison of the agreement on cases with the agreement on referents, the percentages of disagreement differed greatly, while the kappa values were similar. This seemingly inconsistent finding was merely a reflection of the fact that the percentage of disagreement is only a crude measure for agreement that is confounded by coincidental agreement. The a priori chance for referents to be judged fit was relatively high, and this high level lowered the chance of disagreement. Thus, notwithstanding the difference in the percentages of disagreement, the agreement on cases can therefore be considered equally as poor as the agreement on the referents.

The high percentages of false positives and false negatives found in this study appear to have strong implications for the validity of the identification of applicants as being medically fit or unfit for a certain job. First, in this study population, the applicants considered as unfit within each job category by the GOS represented a random sample of all applicants considered unfit by the GOS in that job category. Therefore, the false positive percentages had a direct bearing on the population of all applicants for the job categories involved. About 31% of the applicants considered to be unfit by some physicians may be judged as fit by others if these judgments are based upon summarized data.

Second, it seems reasonable to assume that the group of applicants for whom the GOS physicians did not note a medical diagnosis (70 to 80% of all applicants) would contain only very few false negatives. As a corollary, the false negative percentages within each job category in this study can be estimated to be lower by a factor of perhaps 4—5 in the population of all applicants judged to be fit in that job category. Nevertheless, even if the false negative percentages were actually as low as 0.2 to 2.0%, these percentages would still imply a considerable number of false negatives. The number of false negatives might be comparable with, or even larger than, the actual number of applicants judged unfit. Hence, under these circumstances, the sensitivity of identifying medically unfit applicants with the use of a preemployment medical examination is 50% or less. Moreover, both the false positive percentages and the false negative percentages might in fact be even higher, since the doubtful cases were omitted from the calculations of these percentages.

Considering the practical experience (ie, many of the applicants who were initially judged to be unfit successfully appealed against this GOS judgment), it seems unlikely that closer agreement would have been reached in the present study, even if there had been personal contact between the physician and the applicant.

In this study we did not assess the validity of preemployment medical examinations in identifying the groups of applicants that should be judged medically fit or unfit for certain jobs. Furthermore, it should be clear that this study does not allow conclusions to be drawn about the validity of all preemployment medical examinations of applicants for every job in every country. However, the finding of a lack of consensus among experienced occupational physicians with regard to assessing medical fitness for specific jobs suggests that in many cases the validity of this judgment may be in doubt, even where detailed criteria for fitness are available, because good agreement is a prerequisite of high validity.

Acknowledgments

This study was supported by a grant from the Dutch Ministry of Social Affairs and Employment.

The authors wish to thank Dr LG Fransman for his ongoing support, Dr JJ Radder and Dr MM van der Klaauw for their statistical advice, and the panel physicians for their energy and endurance.

References

1. Cowell JWF. Guidelines for fitness-to-work examinations. Can Med Assoc J 1986;135:985—8.
2. Health and Safety Executive. Pre-employment health screening. London: Health and Safety Executive, 1982. (Guidance note MS20.)
3. Hogan JC, Bernacki EJ. Developing job-related preplacement medical examinations. J Occup Med 1981; 23:469—76.
4. Royal Dutch Society for the Promotion of Health Care. Wat mag en moet bij een aanstellingskeuring [What is allowed and should be done in a pre-employment health examination]. Med Contact 1980;35:849—54.
5. Schilling RSF. The role of medical examination in protecting worker health. J Occup Med 1986;28:553—7.
6. Schussler T, Kaminer AJ, Power VL, Pomper IH. The preplacement examination. J Occup Med 1975;17: 254—7.
7. Lourijsen ECMP, Hoolboom H, de Kort WLAM. The pre-employment medical examination in The Netherlands: a descriptive inventory. In: Rantanen J, Lehtinen S. New trends and developments in occupational health services. Amsterdam: Elsevier Science Publishers, BV, 1991:87—92.
8. Atherley G. Human rights versus occupational medicine. Int J Health Serv 1983;13:265—75.
9. Floyd M, Espir MLE. Assessment of medical fitness for employment: the case for a code of practice. Lancet 1986;2:207—9.
10. Hubbard R, Henifin MS. Genetic screening of prospective parents and of workers: some scientific and social issues. Int J Health Serv 1985;15:231—51.
11. Kelman GR. The pre-employment medical examination. Lancet 1985;2:1231—3.
12. Lappé MA. Ethical issues in testing for differential sensitivity to occupational hazards. J Occup Med 1983;25: 797—808.
13. Rang JF. De aanstellingskeuring: een verwaarloosd arbeidsrechtelijk probleem [The pre-employment medical examination: a neglected legislational issue]. Soc Maandbl Argeidsomst 1978;33:315—9.
14. Rothstein MA. Discriminatory aspects of medical screening. J Occup Med 1986;28:924—9.
15. Sergeant H. Pre-employment psychiatric examinations. Lancet 1984;2:212—4.
16. Todd JW. Pre-employment medical examination. Lancet 1965;1:797—9.
17. Alexander RW, Brennan JC, Maida AS, Walker RJ. The value of preplacement medical examinations for nonhazardous light duty work. J Occup Med 1977;19:107—12.
18. Brownlie L, Brown S, Diewert G, Good P, Holman G, Laue S, Banister E. Cost-effective selection of fire fighter recruits. Med Sci Sport 1985;17:661—6.
19. Newill CA, Evans R, Khoury MJ. Pre-employment screening for allergy laboratory animals: epidemiologic evaluation of its potential usefulness. J Occup Med 1986;28:1158—64.
20. de Kort WLAM, Fransman LG, van Dijk FJH. Preemployment medical examinations in a large occupational health service. Scand J Work Environ Health 1991;17: 392—7.
21. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213—20.
22. Fleiss JL. Statistical methods for rates and proportions. New York, NY: John Wiley & Sons, 1972:143—7.
23. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia, PA: WB Saunders Company, 1985.
24. Dixon WJ, Brown MB, Engelman L, Hill MA, Jennrich RI. BMDP statistical software manual. Los Angeles, CA: University of California Press, 1988.

Received for publication: 21 October 1991