Papers

Annual league tables of mortality in neonatal intensive care units: longitudinal study

Gareth J Parry, Craig R Gould, Chris J McCabe, William O Tarnow-Mordi on behalf of the International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group

Abstract

Objective: To assess whether crude league tables of mortality and league tables of risk adjusted mortality accurately reflect the performance of hospitals.

Design: Longitudinal study of mortality occurring in hospital.

Setting: 9 neonatal intensive care units in the United Kingdom.

Subjects: 2671 very low birth weight or preterm infants admitted to neonatal intensive care units between 1988 and 1994.

Main outcome measures: Crude hospital mortality and hospital mortality adjusted using the clinical risk index for babies (CRIB) score.

Results: Hospitals had wide and overlapping confidence intervals when ranked by mortality in annual league tables; this made it impossible to discriminate between hospitals reliably. In most years there was no significant difference between hospitals, only random variation. The apparent performance of individual hospitals fluctuated substantially from year to year.

Conclusions: Annual league tables are not reliable indicators of performance or best practice; they do not reflect consistent differences between hospitals. Any action prompted by the annual league tables would have been equally likely to have been beneficial, detrimental, or irrelevant. Mortality should be compared between groups of hospitals using specific criteria—such as differences in the volume of patients, staffing policy, training of staff, or aspects of clinical practice—after adjusting for risk. This will produce more reliable estimates with narrower confidence intervals, and more reliable and rapid conclusions.

Introduction

Publication of the United Kingdom's patient's charter[1] has led to an increase in the direct comparisons of institutional performance using league tables.[2,3] The principle behind league tables, as formulated by the Department of Health, is that "performances in the public sector should be measured, and that the public have a right to know how their services are performing.

…Publication [of league tables] acts as a lever for change and a spur to better performance."[4]

The British government is committed to using league tables of hospital outcomes as a method of monitoring services and the implementation of best practices.[5] For league tables to be used in comparing performance they must discriminate reliably and rapidly between hospitals that perform well and those that perform poorly. To act as levers for effective change, differences identified by league tables must be sufficiently stable or definitive to represent a credible argument for change.

We studied mortality in nine neonatal intensive care units over 6 years. Earlier work has shown that comparisons of hospital mortality may be unreliable unless adjustments are made for clinical risk and the severity of illness.[6] We adjusted for these factors using the clinical risk index for babies (CRIB) score.[7,8] We wanted to test whether ranking hospitals by their crude or risk adjusted mortality was reliable in indicating performance.

Subjects and methods

The cohort comprised infants younger than 31 weeks' gestation or weighing less than 1500 g at birth who were admitted to one of nine neonatal intensive care units in the United Kingdom between 1988 and 1994. Gestation, birth weight, congenital malformations, and routine physiological data from the first 12 hours after birth were abstracted from clinical notes by trained researchers. Congenital malformations which were not inevitably lethal were scored according to the clinical risk index for babies as either acutely life threatening or not acutely life threatening by a consultant paediatrician (WOT-M) who was unaware of which hospital treated the infant. The clinical risk index for babies score ranges from 0 to 23; higher scores indicate increasing clinical risk and severity of illness.

All infants admitted to neonatal intensive care units were included in the analysis unless they had an inevitably lethal congenital malformation or were more than 28 days old. Infants who died before 12 hours of age had clinical risk index for babies scores calculated from physiological data recorded up until the time of death. The outcomes of infants who had been

Medical Care Research Unit, School of Health and Related Research, University of Sheffield, Sheffield S3 7XL

Gareth J Parry, *research fellow in health services research*

Sheffield Health Economics Group, School of Health and Related Research

Chris J McCabe, *lecturer in health economics*

Department of Child Health, Ninewells Hospital and Medical School, Dundee DD1 9SY

Craig R Gould, *data manager*

William O Tarnow-Mordi, *reader in neonatal medicine and perinatal epidemiology*

Correspondence to: Mr Parry g.parry@sheffield. ac.uk

Table 1 Number of infants seen, mean (SD) birth weight (g), mean (SD) gestation (weeks), and median (25th, 75th centile) score on the clinical risk index for babies (CRIB) by hospital and number of infants seen in nine neonatal intensive care units. The clinical risk index for babies score ranges from 0 to 23; higher scores indicate an increased risk of hospital mortality. The total number of infants seen over the six years of the study was 2671

Hospital (No infants seen)	Birth weight	Gestation	Clinical risk index for babies (CRIB) score
1 (610)	1132 (311)	30 (3)	3 (1, 7)
2 (289)	1200 (257)	29 (3)	2 (1, 4)
3 (338)	1197 (342)	29 (3)	2 (1, 5)
4 (389)	1157 (283)	29 (3)	3 (1, 7)
5 (475)	1168 (281)	29 (3)	2 (1, 7)
6 (111)	1223 (232)	29 (3)	2 (1, 5)
7 (123)	1169 (278)	29 (3)	4 (1, 6)
8 (66)	1170 (271)	29 (3)	4 (1, 7)
9 (270)	1132 (299)	29 (3)	3 (1, 8)

Table 2 Proportion of annual number of deaths in nine neonatal intensive care units to number of admissions by hospital and year and for all years combined

Hospital	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	All years
1	16/83	25/82	6/89	22/113	16/127	21/116	106/610
2	6/37	7/51	9/53	12/49	5/43	8/56	47/289
3	6/42	10/49	15/62	5/68	10/53	14/64	60/338
4	16/54	17/65	9/68	15/65	17/72	15/65	89/389
5	19/75	12/62	20/100	12/84	10/78	15/76	88/475
6	5/24	6/22	4/24	1/13	1/17	0/11	17/111
7	8/22	7/29	1/22	2/18	7/20	3/12	28/123
8*	4/13	3/20	6/18	3/15	—	—	16/66
9	14/39	10/44	17/54	10/45	14/39	11/49	76/270
Total	94/389	97/424	87/490	82/470	80/449	87/449	527/2671

*Data were not available for hospital 8 for years 5 and 6.

transferred between hospitals were attributed to the hospital that had provided most of the care between 12 and 72 hours after birth. Infants who died in the labour ward or in transit, before admission to a neonatal intensive care unit, were excluded from the study. Mortality was defined as death occurring in any hospital before the infant was discharged. The validity of the score in adjusting for risk was assessed using the area under the curve of the receiver operating characteristic[9] and the Hosmer-Lemeshow goodness of fit test.[10]

Crude league tables were formed without adjusting for case mix. League tables of risk adjusted mortality took account of the infant's initial risk of mortality in each hospital using the clinical risk index for babies score. In both crude and risk adjusted league tables the difference between observed mortality and expected mortality for every 100 infants admitted to the hospital (W score) was used as an indicator of hospital performance.[11] (Information on the W score can be found in the appendix.) A negative W score indicated that a hospital had mortality that was lower than expected, and a positive W score indicated that a hospital had higher than expected mortality. Hospitals were ranked for each year according to their W score. The mortality of the whole cohort was applied to each hospital to give the expected number of deaths for the crude league tables. For the risk adjusted league tables the values for expected mortality came from a logistic regression model that related the score on the clinical risk index for babies to hospital mortality in the whole cohort.

To compare performance it was necessary to determine whether outcomes were significantly different between hospitals and which hospitals performed significantly better or worse than expected. Logistic regression was used in both the crude and adjusted tables to indicate whether significant differences existed in hospital mortality by including a term for hospital. For the risk adjusted league tables, the clinical risk index for babies score was included before the term for hospital in the regression. In years where the term for hospital was significant, hospitals embodying best practice were identified as those where W scores had 95% confidence intervals wholly less than 0.[11] Hospitals where W scores had 95% confidence intervals wholly more than 0 were identified as performing poorly.

To test whether performance at individual hospitals was consistent over time, W scores were included in a two way analysis of variance with random time and hospital effects.[12] Performance was consistent if the variation of W scores between hospitals was greater than variation within a hospital. Hospitals were numbered according to their rankings in the first of the risk adjusted league tables.

Results

Data were obtained for 2671 infants admitted between 1 July 1988 and 30 June 1994. Infants were assigned to six annual groups beginning in July 1988. No data were available for hospital 8 in years 5 and 6. Mean birth weight was 1164 (SD 295) g, and mean gestation was 29 (SD 3) weeks (table 1). The total number of annual admissions for all hospitals combined ranged from a minimum of 389 in year 1 to a maximum of 490 in year 3 (table 2). Crude mortality for all hospitals in the whole cohort was 19.7% (527/2671). For all years crude hospital mortality ranged from a low of 15.3% (17/111) for hospital 6 to a high of 28.1% (76/270) for hospital 9.

League tables of crude mortality

The overall mortality rate of 19.7% was assigned as the expected mortality in calculating W scores for the league tables of crude mortality (fig 1). The term for hospital was significantly associated with mortality only in years 3 and 5. In years 1, 3, 4, and 5 at least one hospital performed either worse or better than expected. In the random effects model the term for hospital was significant in explaining crude W scores over the whole period ($F_{8,43} = 2.164$, $P = 0.0498$). Variation between hospitals accounted for only 17% (95% confidence interval 0% to 42%) of the total variation. Most of the variation in crude mortality was accounted for within, but not between, hospitals. Crude W scores for each hospital were unstable over time. When all years were combined the term for hospital was significantly associated with mortality, and hospital 9 performed worse than expected.

Risk adjusted league tables

When the clinical risk index for babies score was fitted to all infants seen over the six year period the logistic regression model gave the probability of mortality (p) as:

$$p = \frac{e^G}{1 + e^G}$$

where $G = -3.492 + 0.372 \times CRIB$ (clinical risk index for babies) score and e = exponential constant.

The model gave an area under the curve of the receiver operating characteristic of 0.87 (SE 0.02). The Hosmer-Lemeshow goodness of fit test gave a χ^2 value of 15.7 (df = 9, P = 0.074) providing evidence of a satisfactory fit.

In the annual risk adjusted league tables the 95% confidence intervals were wide and overlapping (fig 2). Risk adjusted mortality differed between hospitals only for year 1, when hospital 9 bordered on higher than expected hospital mortality, and year 3, when hospital 1 had lower than expected mortality. The 95% confidence intervals for the apparently best performing and worst performing hospitals overlapped in every year except year 3. When all years were combined there was significant variation in risk adjusted outcomes between hospitals, and hospital 1 had lower than expected risk adjusted hospital mortality.

In the random effects model the term for hospital was not significant in explaining the adjusted W scores over the six years ($F_{8,43} = 0.95$, P = 0.487). The variation between hospitals accounted for 0% (0% to 21%) of the total variation in risk adjusted mortality. Virtually all variation in adjusted W scores was accounted for within, not between, hospitals. For each hospital, adjusted W scores varied considerably over time.

Discussion

Problems with annual league tables

There are three fundamental problems with compiling annual league tables of the performance of individual hospitals. The first problem is the need to make accurate adjustments for differences in case mix. The second is the uncertainty that occurs no matter how accurate the adjustment for case mix when estimates of outcome are made for hospitals that treat relatively small numbers of patients each year.[6 7 13 14] The third problem is the lack of consistency in the apparent performance of hospitals over time. Goldstein and Spiegelhalter have described the problems of accounting for case mix and estimates of outcome in hospitals that see small numbers of patients[15]; our study confirms their results and illustrates the inconsistency in the apparent performance of hospitals over time.

In our study, the crude annual mortality tables (fig 1) had no resemblance to the tables that were adjusted for risk (fig 2); this highlights the necessity of risk adjustment. After adjusting for risk, however, the differences in any particular year were not sufficiently robust that they would allow the hospitals that were ranked the highest to be used as models for the others because of the uncertainty inherent in estimates of outcome in these small samples of patients.[13 15 16] Excluding smaller hospitals from annual league tables would produce narrower confidence intervals, but would be arbitrary and undermine attempts to make hospitals accountable.

The multiple comparisons performed in league tables increase the risk of finding differences due to chance. This could be reduced by using a more conservative significance level, such as P = 0.01.[12] However, this would increase the confidence intervals and further reduce the ability to discriminate reliably between hospitals.
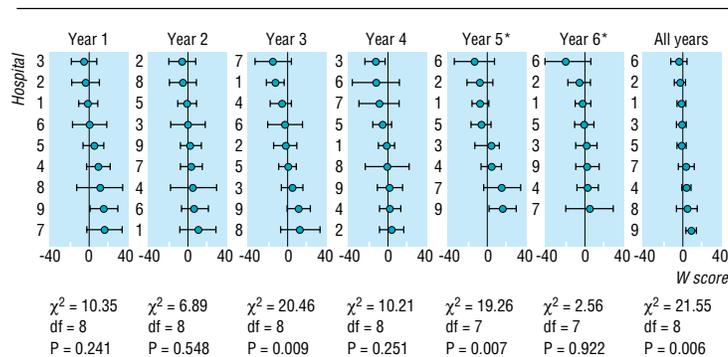
Fig 1 Annual league tables of crude mortality for nine neonatal intensive care units for each year and for all years combined. Mortality is lower than expected when the W score and 95% confidence interval are both <0. Hospitals were ranked in descending order of apparent performance by W score
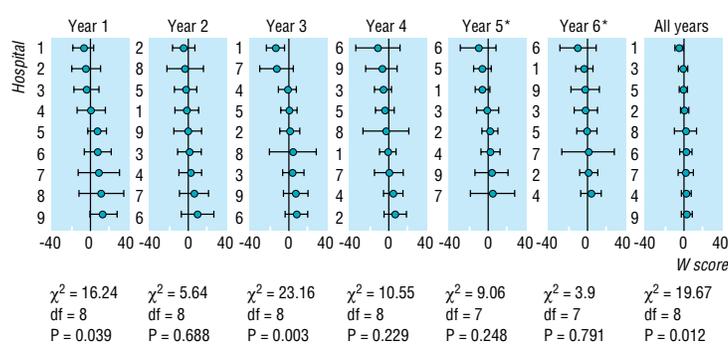
Fig 2 Annual league tables of risk adjusted mortality in nine neonatal intensive care units for each year and for all years combined. Mortality is lower than expected when the W score and 95% confidence interval are both <0. Hospitals were ranked in descending order of apparent performance by W score. The 95% confidence intervals of the hospitals that apparently performed best and those that apparently performed worst overlapped for every year except year 3

Rather than attempt to obtain precise estimates of performance it may be better to determine if a hospital has exceeded a certain standard. Comparison of a hospital with an annual central measure, or average, would result in around half of the hospitals performing below standard. Alternatively, if it was decided that good or bad performance was indicated by a hospital having a W score outside two standard errors of 0 (P = 0.05) then one hospital in 20 would be identified by chance. It is impossible to assess whether a single significant result reflects a difference in the quality of care or chance. Only results consistently outside these cut off points should be used to indicate good or bad performance.

The finely designated rankings that are implied by annual league tables cannot be justified. Publication of the tables is likely to encourage doubts over whether optimal care has been delivered. The resulting anxiety, stigma, and guilt among parents and providers of the service will be unnecessary and mostly unfounded.

Overall, hospital 1 did perform significantly better than expected. It is debatable whether this makes it a model hospital since its performance was inconsistent. It took 6 years for hospital 1 to be identified as the most

successful hospital. By then the result may only have been of historical interest. One argument for using risk adjusted league tables is that the feedback they provide stimulates improved performance in the year following what was ranked as a year of supposedly poor performance. Proponents of this view might argue that this occurred for hospitals 6 and 9 between years 3 and 4 (fig 2). However, feedback cannot explain these changes in ranking since in this study hospitals did not receive feedback until after year 6. Furthermore, the random effects model established that variation in ranking from year to year was explained by random variation within hospitals rather than systematic differences between them. The most likely explanation for the apparent improvement of hospitals 6 and 9 and the decline of hospital 1 between years 3 and 4 is random variation resulting in regression towards the mean.[12] That is, low ranking hospitals will improve by chance alone, which can mistakenly be attributed to some action introduced in an attempt to improve performance. In fact any action prompted by these league tables would have been equally likely to have been beneficial, irrelevant, or detrimental.

Alternatives to annual league tables

Instead of outcomes, it may be more useful to compare the implementation of measures of process which have been proved to be accurate in randomised controlled trials. Mant and Hicks showed that comparing the use of proved treatments for myocardial infarction would substantially reduce the number of patients and the time required to identify significant differences in hospital performance.[17] However, comparisons of the proportions of patients treated are only valid if patients in different hospitals are equally eligible to receive the treatments under comparison at all hospitals.[18]

Even if it is possible to find hospitals that consistently perform better than others it may not be obvious why they perform better.[19] The efficient use of resources requires that the specific characteristics of best practice are identified, so that no unnecessary changes are made. Perhaps the best use of league tables of outcomes would be to help formulate hypotheses about characteristics of best practice for further investigation.[20] Rather than relying on league tables to monitor performance, it would be better to undertake prospective studies of outcome in large groups of hospitals and test whether improvements in outcome are associated with differences in specific characteristics—such as the volume of patients, the staffing levels, or the training and expertise of staff—after adjusting for risk.[21–23] This approach, which has been adopted in the UK Neonatal Staffing Study,[22 23] would identify more reliably and rapidly the organisational characteristics likely to improve outcome. This would allow institutions to become more accountable to the public by showing that their policies are based on reliable evidence. Our findings may apply to other areas where the use of mortality league tables are being considered,[5] such as the ranking of individual hospitals by annual rates of infection, surgical complications, or surgical mortality, Given that league tables did not seem effective in evaluating neonatal intensive care units, those who support the use of mortality league tables must now show why they might be useful elsewhere.[5 25]

Key messages

- League tables are being used increasingly to evaluate hospital performance in the United Kingdom
- In annual league tables the rankings of nine neonatal intensive care units in different hospitals had wide and overlapping confidence intervals and their rankings fluctuated substantially over six years
- Annual league tables of hospital mortality were inherently unreliable for comparing hospital performance or for indicating best practices
- The UK government's commitment to using annual league tables of outcomes such as mortality to monitor services and the spread of best practices should be reconsidered
- Prospective studies of risk adjusted outcome in hospitals grouped according to specific characteristics would provide better information and be a better use of resources

Appendix: The W score

The use of the W score has become established in the presentation of comparisons between observed and expected mortality after trauma.[11] In this study the W score is calculated as follows:

W = 100 × (observed − expected deaths)/No of admissions.

For the crude tables, expected deaths are 19.7% of the annual number of admissions in each hospital. For the adjusted tables, expected deaths are predicted by a logistic regression model based on the score on the clinical risk index for babies.

The 95% confidence interval for the W score is given as: W ± 1.96 SE (W), where

$$SE(W) = \frac{100}{n} \sqrt{\Sigma p_i (1 - p_i)}$$

n = No of admissions; Σ implies the sum over all infants in each hospital in each year; and p_i = probability of death in hospital for infant i.

For infant i in the crude tables: $p_i = 0.197$. For infant i with a score on the clinical risk index for babies (CRIB) of $CRIB_i$, in the adjusted tables:

$$p_i = \frac{e^{G_i}}{1 + e^{G_i}}$$

where $G_i = -3.492 + 0.372 * CRIB_i$ and e = exponential constant.

1 Department of Health. *Patient's charter*. London: HMSO, 1991.
2 Clinical Outcomes Working Group. *Clinical outcome indicators*. Edinburgh: Scottish Office, 1994.

3   McKee M. Indicators of clinical performance. *BMJ* 1997;315:142.
4   Butler R. Discussion. In: Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statis Soc (A)* 1996;159:409.
5   Wise J. Clinical indicators for hospitals announced. *BMJ* 1997;315:76.
6   Tarnow-Mordi WO, Ogston AR, Wilkinson AR, Reid E, Gregory J, Saeed M, et al. Predicting death from initial disease severity in very low birthweight infants: a method for comparing the performance of neonatal units. *BMJ* 1990;300:1611-4.
7   The International Neonatal Network. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care hospitals. *Lancet* 1993;342:192-8.
8   De Courcy Wheeler RHB, Wolfe CDA, Fitzgerald A, Spencer M, Goodman JD, et al. Use of the CRIB (clinical risk index for babies) score in prediction of neonatal mortality and morbidity. *Arch Dis Child* 1995;73:F32-36.
9   Hanley JA, McNeil BJ. A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839-43.
10 Hosmer DW, Lemeshow S. *Applied logistic regression.* New York: Wiley, 1989.
11 Yates DW, Woodford M, Hollis S. Preliminary analysis of the care of injured patients in 33 British hospitals: first report of the United Kingdom major trauma outcome study. *BMJ* 1992;305:737-40.
12 Armitage P, Berry G. *Statistical methods in medical research.* 3rd ed. London: Blackwell Scientific, 1994.
13 Rockall TA, Logan RFA, Devlin HB, Northfield TC, for the national audit of acute upper gastrointestinal haemorrhage. Selection of patients for early discharge or outpatient care after acute upper gastrointestinal haemorrhage. *Lancet* 1995;346:346-50.
14 Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York state's approach. *N Engl J Med* 1995;332:1229-32.
15 Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statist Soc (A)* 1996;159:385-443.

16 Tarnow-Mordi WO, Parry GJ, Gould C, Fowlie PW. CRIB and performance indicators for neonatal intensive care units (NICUs) [letter]. *Arch Dis Child* 1996;74:F79-80
17 Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ* 1995;311:793-6.
18 Scottish Neonatal Consultants Collaborative Study Group and International Neonatal Network. Trends and variations in use of antenatal corticosteroids to prevent neonatal respiratory distress syndrome: recommendations for national and international comparative audit. *Br J Obstet Gynaecol* 1996;103:534-40.
19 Pearson G, Shann F, Barry P, Vyas J, Thomas D, Powell C, et al. Should paediatric intensive care be centralised? Trent versus Victoria. *Lancet* 1997;349:1213-7.
20 Kendrick S. Discussion. In: Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statist Soc (A)* 1996;159:433.
21 Pollack MM, Cuerdon T, Patel KM, Ruttimann UE, Getson PR, Levetown M. Impact of quality of care factors on pediatric intensive care unit mortality. *JAMA* 1994;272:941-6.
22 Tarnow-Mordi WO, Tucker JS, McCabe CJ, Nicolson P, Parry GJ for the UK Neonatal Staffing Study Collaboration Group. The UK neonatal staffing study: a prospective evaluation of neonatal intensive care in the United Kingdom. *Semin Neonatology* 1997;2:171-9.
23 Tarnow-Mordi W, Parry G. Inappropriate and appropriate comparisons of intensive care units using the clinical risk index for babies and pediatric risk of mortality scores. In: Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statist Soc (A)* 1996;159:436.
24 The UK Neonatal Staffing Study. http://www.child-health.dundee.ac.uk/research/UKneonatal-staffing/.
25 Fowlie PW, Gould CR, Tarnow-Mordi WO, Parry GJ, Phillips G. CRIB (clinical risk index for babies) in relation to nosocomial bacteraemia in very low birthweight or preterm infants. *Arch Dis Child* 1996;75:F49-52.

(Accepted 12 February 1998)

Effect of flutamide on survival in patients with pancreatic cancer: results of a prospective, randomised, double blind, placebo controlled trial

Brian A Greenway

Abstract

Objectives: To assess whether flutamide (Drogenil), a pure androgen receptor blocking agent, improves survival in patients with pancreatic carcinoma and thus whether testosterone is a major growth factor for this tumour.

Design: A prospective, randomised, double blind placebo controlled trial.

Subjects: 49 patients with a clinical diagnosis of pancreatic carcinoma.

Interventions: 24 patients received flutamide and 25 received placebo.

Main outcome measures: Death of the patient.

Results: Analysis of all patients at 6 months and 1 year showed 14 and eight patients alive, respectively, in the flutamide group compared with 10 and one in the placebo group. After exclusion of those patients in both groups who received less than 6 weeks' treatment because of advanced disease and early death the comparable results were 14 (88%) and eight (50%) alive in the flutamide group compared with 10 (50%) and one (5%) in the placebo group. Median survival for all patients was 8 months in the flutamide group compared with 4 months in the placebo group. With the 6 week exclusions median survival was 12 months compared with 5 months, respectively.

Conclusions: This study supports the concept that testosterone is a growth factor for pancreatic carcinoma and that blockade of androgen receptors offers an appropriate new approach to treatment.

Introduction

Pancreatic cancer is difficult to diagnose and unsatisfactory to treat, with most patients dying within 6 months of diagnosis and virtually all by 1 year.[1 2] At present surgery offers the only prospect of longer survival.

Testosterone may have a positive effect on the growth of pancreatic carcinoma. This would be supported if specific androgen receptor blockade improved survival. The concept is supported by the presence of androgen receptors within human pancreatic cancer tissue,[3] together with the enzymes, aromatase and 5α-reductase, which convert testosterone into either oestradiol or a more active androgen, 5α-dihydrotestosterone, respectively.[4] Furthermore, all patients with pancreatic cancer have low serum testosterone concentrations.[5 6]

Confirmatory evidence for this central role of testosterone came with the demonstration of its growth potentiating action on human pancreatic adencocarcinoma xenografts grown in nude mice, together with the inhibiting action of an antiandrogen.[7]

Patients and methods

The trial was conducted at Hinchingbrooke Hospital, Huntingdon, between October 1993 and January 1997

Department of Surgery, Hinchingbrooke Hospital, Huntingdon, Cambridge PE18 8NT

Brian A Greenway, *consultant surgeon*