Waiting time information services: how well do different statistics forecast a patient's wait?

DAVID A. CROMWELL AND DAVID A. GRIFFITHS

David A. Cromwell is a Senior Research Fellow at the Centre for Health Service Development, University of Wollongong, Wollongong, NSW.
David A. Griffiths is a Professor at the School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW.

Abstract

This study investigates how accurately the waiting times of patients about to join a waiting list are predicted by the types of statistics disseminated via web-based waiting time information services. Data were collected at a public hospital in Sydney, Australia, on elective surgery activity and waiting list behaviour from July 1995 to June 1998. The data covered 46 surgeons in 10 surgical specialties. The accuracy of the tested statistics varied greatly, being affected more by the characteristics and behaviour of a surgeon's waiting list than by how the statistics were derived. For those surgeons whose waiting times were often over six months, commonly used statistics can be very poor at forecasting patient waiting times.

The use of waiting time statistics to inform referral decisions

A recent analysis of waiting list behaviour at one Australian hospital raised questions about whether patients should use waiting time information services to make inferences about their likely waiting time, or about at which hospital they might have a shorter wait (Cromwell and Griffiths 2002). These questions arose because of two principal issues. The first concerned the extent to which different levels of data aggregation might lead to disseminated statistics being biased by, or unresponsive to, changes in behaviour. The second concerned how potentially small sample sizes might affect the accuracy and reliability of the statistics. A review of waiting time information services had concluded that neither issue could be discounted. The services differed in how their statistics were derived and the services provided little guidance to users on how to appropriately interpret the presented information, even though the statistics could be used in various ways (Cromwell et al. 2002).

That services suffered from these problems may stem from the limited practical advice on waiting time statistics in the literature. Arguments have been made for the use of specific types of data (Don et al. 1987; Mordue et al. 1989), for particular measures, such as the median rather than the mean (Mason 1976; Black 1998; Armstrong 2000) and for how statistics should be presented (White 1980; Pugh 1987). However, there have been no empirical studies that suggest whether waiting time statistics are sufficiently accurate for doctors or patients either to make inferences about expected waiting times or to select surgical units at which waiting times will be shorter. Consequently, a study was undertaken to assess the relative performance of commonly used waiting time statistics in predicting the waiting times of patients.

Method

Waiting list data that provided information on elective surgical activity between 1 July 1995 and 30 June 1998 were collected from a teaching hospital in Sydney, Australia. De-identified data were extracted on all patients admitted or removed from the list during this three year period, together with data on the patients still waiting on 30/6/98. The data covered 46 surgeons in 10 specialties who were active throughout this period. The data set is described in more detail elsewhere (Cromwell and Griffiths 2002).

The evaluation mimicked the operation of a waiting time information service. Services were assumed to disseminate statistics derived from data available at the end of one period (say t) such that they were available to patients who joined the waiting list in the next period (t+1). It was assumed that statistics were updated each month, and that there was no delay in the data becoming available.

The analysis was limited to testing statistics for non-urgent patients, i.e. those assigned to the New South Wales urgency category 8 (denoted here as U8). As the waiting times of patients who change and do not change urgency category have been found to differ (Cromwell and Griffiths 2002), the evaluation was limited to examining how well the statistics forecast the wait of those patients that had not changed category or listing status (e.g. been deferred). For the same reason, the tested statistics were also derived from these patients. Thus, waiting time was simply the total time spent on a list prior to leaving it, or up to the census date, and calculated in days.

The accuracy of the waiting time statistics was measured by calculating the difference between the forecast waiting time and the waiting times of individual patients, and summarising these using the mean square error (MSE) and the mean absolute error (MAE) (Makridakis et al. 1998). Both were used so that the measure of performance did not favour particular statistics (the mean of a random variable will minimise the MSE, while the median will minimise the MAE (DeGroot 1986)). Other assessment criteria were the proportion of patients who waited beyond the forecast time, and the proportion of patients whose wait exceeded that forecast by 90 days or more. Differences between the forecasts produced by various tested statistics were summarised using the mean absolute difference (MAD) (Makridakis et al. 1998).

For most surgeons, the evaluation began from the fourth month of the three year period. This was because data from the first three months were required to derive the initial values of the tested statistics. The exceptions were surgeons Dr012 and Dr026. Here, the evaluation began from the seventh month because activity in the preceding months did not allow one or more types of statistic to be derived.

The forecast waiting time should be compared to the wait of each patient who joins in a particular month. Consequently, the primary rule for defining the last month used in the evaluation was the point at which one or more patients who joined the waiting list in the same month were still on the waiting list (or the end of the time series if no patients were waiting). However, for some surgeons, this defined an end point toward the middle of the three year period, which substantially limited the data available for analysis. It was therefore decided to include some of the patients who were still waiting, and to treat these patients as if they had been admitted. To limit the bias this would cause, only patients already waiting an unusually long time were included. The precise inclusion rules for these patients were defined as follows:

- if the 3rd quartile of the cross-sectional waiting time distribution was between 12 and 18 months, the analysis included any patient still on the list who had waited for more than 18 months;
- if the 3rd quartile was less than 12 months, the analysis included any patient still on the waiting list if (1) they had waited longer than 12 months, or (2) all patients who joined the waiting list in the two following months had left the list.

Table 1 gives the number of months included in the analysis for each surgeon, the number of patients against which the forecasts were compared, and the mean of their overall waiting time distribution. The sample included 240 patients with censored waiting times, spread across 19 surgeons. Surgeon Dr003 contained the most patients included in the data for any one surgeon, namely 37 (20%). The waiting time of these patients with censored waiting times was defined as the time spent on the list up to 1 July 1998, the first day after the data collection period.

Table 1. Number of months, and the number of patients, included in the analysis by surgeon. The patients' average waiting time (days) is also given.

Surgeon	Months in analysis	No. of Patients	Average wait	Surgeon	Months in analysis	No. of Patients	Average wait
Dr001	33	15	23	Dr024	21	85	270
Dr002	32	163	23	Dr025	25	46	135
Dr003	21	183	251	Dr026	29	38	22
Dr004	29	265	86	Dr027	33	116	28
Dr005	29	207	25	Dr028	28	122	21
Dr006	26	303	131	Dr029	21	100	258
Dr007	21	598	86	Dr030	18	77	293
Dr008	28	733	59	Dr031	15	62	>300
Dr009	25	174	65	Dr032	21	238	238
Dr010	33	134	34	Dr033	15	79	>300
Dr011	29	51	46	Dr034	21	57	293
Dr012	26	32	36	Dr035	21	178	236
Dr013	29	556	28	Dr036	21	68	124
Dr014	26	445	29	Dr037	16	33	>300
Dr015	25	202	123	Dr038	33	46	23
Dr016	31	50	72	Dr039	26	137	91
Dr017	21	138	135	Dr040	33	56	15
Dr018	25	51	114	Dr041	28	168	91
Dr019	21	125	118	Dr042	33	47	64
Dr020	24	112	192	Dr043	32	77	62
Dr021	30	77	103	Dr044	15	43	>300
Dr022	25	62	116	Dr045	21	141	236
Dr023	30	193	30	Dr046	21	110	274

Waiting time statistics (forecast functions) that were tested

Because many commonly used waiting time statistics are defined in different ways (Cromwell et al. 2002), the study assessed an array of functions. The tested functions were derived from data of patients admitted during a defined interval (throughput data) and from data of patients still on the waiting list at specific dates (census data). Two sets of statistics were produced from both types of data. The first set was based on surgeon level data, and resulted in each surgeon having a unique series of forecasts. The second set was derived from specialty level data, with surgeons in the same specialty each using the same series.

Within each set, seven forecast functions were defined using a variety of simple smoothing methods. The simplest functions were the mean (MA1) and the median (MD1) waiting time of data collected over one unit of time. For throughput data, a unit corresponded to data collected during one month, while for census data, it corresponded to data collected at one census point. The other forecast functions were:

- the mean (MA3) and the median (MD3) waiting time of throughput data collected during three months, or census data collected on three census dates;
- a 3-month moving average of the MD1 forecast (MA3MD1);
- a 3-month moving average of the MD3 forecast (MA3MD3); and
- an exponentially weighted moving average of the MA1 forecast (EWMA1).

If for some reason a value could not be derived (for example, because there were no admissions during the period), the forecast waiting time was assumed to be that of the preceding period.

The MA3 and MD3 forecasts were based on data aggregated over the last three months in which patients were admitted, or the three most recent census dates on which some patients were on the waiting list. Although unorthodox, this improved the degree of smoothness of consecutive forecasts, and the robustness of the series against the effects of high and low outliers (Cromwell and Griffiths 2002). The only exception to this occurred at the start of the time series if a surgeon did not have three months of data. In this case, the forecast was based on the available data. The MA3MD3 function was also defined slightly differently at the start of the time series. Its first two values were, respectively, the first MD3 value and the average of the first two MD3 values.

The smoothing constant (k) used in the EWMA1 formula ($F_t = k.MA1_t + (1-k).F_t\text{-}1$) was 0.3. This value was chosen because it produced greater smoothing than the other functions (Bissel 1994) and because it represented a good compromise value for most circumstances (Box and Luceno 1998). The initial value for the EWMA1 series was defined to be the initial value of the MA3 function.

Results

This section focuses initially on the accuracy of the forecast functions based on surgeon level data. Their performance varied substantially between surgeons; the variation between the functions themselves was much smaller. Table 2 shows this in terms of the square root of the MSE values (RMSE); the same pattern was observed for the MAE values.

Table 2. Differences in accuracy between the forecast functions based on surgeon level data, stratified by the RMSE (days) of the most accurate function

Least RMSE across	Functions based on throughput data			Functions based on census data		
the functions	No. of surgeons	Minimum range [1]	Maximum range [2]	No. of surgeons	Minimum range	Maximum range
< 30	7	2	11	8	2	15
30 - 59	9	2	26	7	3	24
60 - 89	8	2	12	8	3	13
90 - 119	7	5	14	6	6	13
120 - 149	2	11	13	3	6	10
150 - 179	3	10	23	1	6	6
180 - 209	2	6	36	2	6	8
210 - 239	1	27	27	4	9	10
240 - 269	3	22	58	3	8	24
270 plus	4	32	48	4	11	29

[1] the minimum difference between the smallest and largest RMSE values among the surgeons

[2] the maximum difference between the smallest and largest RMSE values among the surgeons

The performance of the functions decreased as the average waiting time of patients joining the list increased. Indeed, for each function, the correlation between a surgeon's RMSE value and the average waiting time of the patients added to the surgeon's list was high (Pearson's $r > 0.93$). This was because the variation in patient waiting times increased as the average rose. Once the average wait exceeded six months, the RMSE of all functions was typically greater than 150 days. The only exceptions were the functions based on throughput data for surgeon Dr029, but this surgeon had unusually low variation in patient waiting times given the average wait.

The consequence of this variation is more easily interpreted in relation to the '90 day' criterion. For the 13 surgeons with an average wait above six months, the proportion of patients whose wait exceeded that forecast by 90 days

was greater than 40% for the best performing throughput-data function. Similar overall levels of performance were found for the functions derived from census data. For the best performing function to have a proportion below 20%, the average waiting time of patients joining a surgeon's list had to be less than three months.

The proportion of patients who waited beyond the time forecast was affected by how patient waiting times varied. For the functions based on throughput data, those proportions above 50% tended to be associated with surgeons whose average wait was above 3 months, and whose waiting times increased over the data analysis period. An equivalent association was observed for the functions based on census data.

As noted above, the differences between the RMSE values of the seven types of function were small in comparison with the differences between surgeons (see Table 2). The largest differences generally occurred amongst the functions using throughput data. Here, there was an upward trend in the range of RMSE values as the performance of the functions became worse. For those surgeons with a RMSE above 180 days, the range generally exceeded 30 days. The maximum range among functions using census data was less (29 days), and a large range was not limited to surgeons for which functions performed poorly.

To help distinguish any overall pattern, the performance of the functions for a single surgeon were ranked from best (=1) to worst (=7) for each criterion. Table 3 shows the average rank statistics for the functions when derived from both the throughput data and census data. With respect to the throughput data functions, the MA3 function performed consistently better than the other functions. The MA3 function also performed the best of the census data functions, although the more standard function based on data from one census date, the MA1 function, performed similarly well. With respect to the performance of functions based on a mean or median, the analysis suggests that those based on the mean are to be preferred. Functions based on the mean produced forecasts that were typically higher than those produced by the median-based functions. The outcome of this was most visible in relation to the proportion of patients who waited beyond the time forecast. The median-based functions had a higher proportion of patients waiting in excess, and the proportion was generally greater than 50%.

Table 3. Rank statistics summarising the difference between forecast and actual waiting time

Surgeon level, throughput data							
Average Rank	MA1	MD1	MA3	MD3	MA3MD1	MA3MD3	EWMA1
RMSE [1]	4.4	5.3	2.8	4.2	3.7	4.4	3.2
MAE [1]	4.4	4.7	3.8	3.8	3.6	3.8	4.0
%Wt > Fc [2]	3.6	4.3	2.6	4.2	3.7	4.5	3.1
%Wt > Fc+90 [1]	2.4	2.8	2.4	3.5	3.0	3.8	3.1
Surgeon level, census data							
Average Rank	MA1	MD1	MA3	MD3	MA3MD1	MA3MD3	EWMA1
RMSE	2.4	4.7	2.5	5.0	4.7	5.4	3.3
MAE	3.2	4.7	3.3	4.2	4.7	4.6	3.4
%Wt > Fc	2.4	4.6	1.9	4.7	4.1	4.7	2.2
%Wt > Fc+90	1.7	3.2	1.7	3.8	3.9	4.4	2.4

[1] Functions with lower values given lower ranks, 1 being the highest rank, 7 being the lowest

[2] Functions with the proportions closer to 50% given the lower rank, 1 being the closed, 7 being the farthest away

Against the RMSE or MAE criteria, differences in the performance of the functions using the two types of data were limited. The largest differences in the RMSE values (>40 days) were for surgeons Dr029 and Dr030, although for all functions except MA1 and MD1, the function based on throughput data performed best for between 30 to 32 surgeons. The differences between equivalent functions were more apparent in terms of the proportion of patients who waited beyond the forecast time. As the functions' forecasts correspond to a

prediction of the expected waiting time, good performance can be regarded as values near 50%. The proportions resulting from the throughput data statistics were typically distributed around this point, with only a few surgeons having a value for any function outside 30-70%. In contrast, the proportions tended to be higher for the census data functions. For 29 surgeons, over 50% of patients waited beyond their forecast waiting time regardless of the census data function used. The performance of the census data functions against the 90-day criterion was also worse in some instances. For 6 of the 13 surgeons with an average wait above six months, the proportions resulting from the census-data functions were at least 15% higher than the equivalent throughput data functions in most cases.

The analysis of the forecasts derived from data aggregated at a specialty level produced similar patterns of performance from the various functions. The best performing function using throughput data was again the MA3 function, while the MA1 and MA3 functions using census data both performed well. Consequently, the differences between functions derived from surgeon and specialty level data are described in relation to the throughput data MA3 function (MA3(TH)) and the census data MA1 function (MA1(CS)). The MA1(CS) function was selected because census data statistics are commonly based on data from one census date.

Table 4 gives the RMSE values for both functions when derived from surgeon and speciality level data, as well as the mean absolute difference between forecasts when produced from surgeon and specialty level data. With respect to the MA3(TH) function, the effect of using specialty level data was not uniform. For many surgeons, it made little difference, although performance was worse for 30 of the 46 surgeons overall. Only for surgeon Dr031 did the specialty level function produce substantially more accurate forecasts. This was due to an interval during which most admissions were patients with low waiting times, and which produced unrealistically low forecasts. For other surgeons (Dr023, Dr026, and Dr044), performance was noticeably poorer, due to the large difference between the forecasts produced from the two levels of aggregation. That large differences did not always produce poorer performance seemed to be more luck than an inherent quality. For instance, waiting times between the four surgeons in urology differed markedly, and the statistics produced from specialty data were typically too high for three of the surgeons, while being too low for fourth. Yet the higher forecasts were occasionally accurate as the waiting times of patient joining the list fluctuated, and the surgeon level functions performed equally poorly because they lagged behind the fluctuations.

There was also no consistent pattern to which level of aggregated data produced more accurate MA1(CS) forecasts. However, for six of the seven surgeons with the largest differences, the statistics based on specialty level data performed worse. Surgeon Dr036 was the one exception. Here, the specialty level forecasts were more accurate because an increase in waiting times at a specialty level was fortuitously predictive of an increase in waiting times for this surgeon. Where specialty level forecasts performed slightly better for other surgeons (Dr011, Dr013, Dr021 and Dr032), successive forecasts were noticeably smoother than those produced from surgeon level data. However, the percentage improvement in performance was not large.

Examining the other criteria revealed differences between equivalent functions based on surgeon and specialty level data, most notably with respect to the proportion of patients whose wait exceeded that forecast. For the MA3(TH) function, the values were no longer condensed into the middle of the graph, indicating a tendency for under-estimation (a shift higher) or over-estimation (a shift lower). A similar scattering effect was evident in the proportions resulting from the MA1(CS) function, even though the proportions produced by the forecasts based on surgeon level data were already fairly well spread.

With respect to the proportion of patients whose wait exceeded that forecast by 90 days or more, the use of specialty level data reduced the proportion for the majority of surgeons (20 for the MA3(TH) function and 24 for the MA1(CS) function). However, the lower proportions produced by forecasts based on specialty data did not necessarily indicate more accurate predictions, simply higher values than those produced from surgeon level data. When measured using the RMSE criterion, performance was worse in many cases.

Table 4: comparison of forecasts produced from data aggregated at surgeon and specialty level for the MA3(TH) and the MA1(CS) functions

Surgeon	MA3 function, throughput data			MA1 function, census data		
	Surgeon level data (RMSE)	Specialty level data (RMSE)	M.A.D. between forecasts	Surgeon level data (RMSE)	Specialty level data (RMSE)	M.A.D. between forecasts
Dr001	33	27	13	35	50	36
Dr002	21	20	1	37	36	2
Dr003	189	195	33	212	217	11
Dr004	59	74	39	59	77	51
Dr005	42	52	41	68	56	45
Dr006	71	99	70	84	93	26
Dr007	78	85	13	85	80	11
Dr008	57	57	10	59	56	15
Dr009	66	63	13	64	64	16
Dr010	26	31	40	28	34	51
Dr011	60	54	38	75	54	36
Dr012	51	47	51	47	52	44
Dr013	53	57	22	73	59	28
Dr014	38	43	32	39	47	33
Dr015	86	112	31	111	110	15
Dr016	72	73	47	70	69	67
Dr017	102	112	20	112	106	18
Dr018	112	116	19	111	116	17
Dr019	128	122	17	123	115	15
Dr020	207	207	39	207	201	40
Dr021	112	103	27	115	94	36
Dr022	128	126	34	130	124	16
Dr023	24	91	80	28	92	89
Dr024	248	271	42	247	261	20
Dr025	118	124	36	112	116	18
Dr026	20	94	98	24	103	112
Dr027	22	21	6	24	22	12
Dr028	16	17	4	19	17	5
Dr029	115	122	15	177	168	11
Dr030	186	189	26	237	231	16
Dr031	347	298	69	352	360	22
Dr032	188	184	19	213	194	33
Dr033	246	259	37	260	276	28
Dr034	218	222	59	232	234	17
Dr035	199	189	37	195	196	11
Dr036	119	112	94	134	100	95
Dr037	269	289	69	275	303	33
Dr038	40	66	69	74	68	88
Dr039	111	115	19	107	108	4
Dr040	25	50	82	26	62	92
Dr041	89	80	68	81	154	154
Dr042	72	83	80	78	143	150
Dr043	70	97	91	61	158	170
Dr044	336	387	88	340	367	51
Dr045	285	279	21	272	262	18
Dr046	324	324	25	299	304	20

Discussion

The aim of this study was to provide some insight into how well waiting time statistics predict the waiting times of patients joining a waiting list. It was triggered by the differences between waiting time information services in the types of statistics used, and by the lack of empirical evidence about whether waiting time statistics should be used to make inferences about the waiting times of patients or to select surgical units at which waiting times will be shorter.

Although the primary focus of the analysis was to examine the differences between the functions, the most noticeable feature of the analysis was the large variation in the accuracy of the statistics between surgeons. These differences were related to the level of variation in waiting times between patients who joined the list at similar times. This variation increased as the average wait rose, a relationship that arises predominantly from how patients are selected from the waiting list. Variation relative to the average waiting time was least for surgeons who most closely approximated a first come, first served admission policy (e.g. Dr029). Another important factor affecting forecast accuracy was how well a function responded to changes in waiting times over time.

The observed levels of performance have various implications for how services aim to meet the information needs of doctors and patients. If a service aims to provide a prediction of how long a patient might expect to wait for admission, the results suggest that services should not simply disseminate an estimate of an expected waiting time. Doing so would be misleading as the distribution of patient waiting times around the expected value can be large, especially when the average exceeds six months. Some estimate of the time below which the majority of patients wait is also required. Moreover, it suggests services should give a clear statement of the dangers of using the information in this way.

In terms of comparing surgical units, the analysis suggests that it is sufficient to present an estimate of the expected wait. The spread of waiting times increased as the average wait rose. Consequently, if the expected waiting time at one unit is greater than at another, it is likely that the same is true of other measures of location (i.e. the percentiles of the distribution). Nonetheless, the key issue concerns how large the difference should be between average waiting times at two surgical units before patients (deciding where to be referred) can infer that their waiting time will be shorter at the unit with the lower average. Unfortunately, it is difficult to derive an exact value for this difference. The waiting time statistics were not produced from a model which gives a probabilistic prediction interval. Moreover, for the statistics to be accurate, future behaviour must closely approximate that of the recent past, and this may not occur (Cromwell and Griffiths 2002). Still, without guidance, users may read too much into small differences.

A crude "rule of thumb" for the minimum significant difference was derived for the MA3(TH) and MA1(CS) functions using the method described in the appendix. This resulted in the following guidelines:

- services estimating expected waiting times based on throughput data should advise users that, unless average waiting times differ by at least one half of the midpoint between the two averages, they should not choose one unit over another based on waiting time information alone;
- services estimating expected waiting times based on census data should advise users that the average waiting times should differ by at least 60 days (regardless of the size of the averages) before they should choose one unit over another.

Of course, the method assumes that there will be no significant change in average waiting times while a patient is on the waiting list, and this also needs to be clearly stated.

Overall, the results show that users should be cautious about using services to make inferences about their likely waiting times. The services are perhaps best used to assess whether the waiting times of a preferred surgical unit are too long. What constitutes "too long" will vary between patients, and judgements will also depend upon what statistics a service uses. But if a service presents average waiting time statistics, the results suggest that surgical units with an average wait above six months should be avoided.

Attention is now turned to the differences in performance between the statistics tested. The study confirms that functions will produce different numerical forecasts, although how this related to forecast accuracy depended upon various factors. Statistics based on the mean seem to be more accurate than those based on the median. Moreover, for the median-based statistics, the proportion of patients whose wait exceeded the forecast time was generally not near 50%. Thus, in this context, the median may give a more misleading picture than the mean, and suggests general statements about which measure is to be preferred should be discounted.

In situations where surgeons manage their own lists, it seems that statistics based on surgeon level data are to be preferred to those based on specialty level data. Functions using specialty data could perform better due to improved robustness, but the gains were small compared with the poorer performance that could arise due to large differences among surgeons in the same specialty. Moreover, when forecasts based on surgeon level data were much more accurate, it was typically when those functions had a RMSE of less than 90 days. In comparison to these acceptable levels of accuracy, the equivalent specialty-level functions performed poorly. It seems that aggregating data over three months is sufficient to produce reliable statistics from throughput data aggregated by surgeon and urgency category. However, this result was based on throughput data being aggregated over three months in which there were admissions. Using data from one census date would seem sufficient if census data were used.

Finally, the statistics based on throughput data generally performed better than those derived from census data. The main exception to this occurred when a surgeon admitted a series of patients with low waiting times, producing forecasts that severely underestimated the eventual waiting times of most patients. When the census was low, the performance of the census data statistics could be affected by the long waiting times of one or two patients. When the census was long, the incomplete waiting times of many patients appeared of greater importance, as the census data statistic was often too low. However, the amount by which statistics based on throughput and census data differed was greatly influenced by the order in which patients were admitted. For those surgeons with a tendency to select patients from the middle of a list, the two statistics often had similar values.

Nonetheless, the levels of performance did not differ between statistics as much as might have been expected. This was due to several factors. First, the difference between the forecasts were often less than the overall variation in waiting times of patients over the period being analysed, and this made the MSE (and MAE) fairly insensitive measures. Second, for some surgeons, admissions were fewer over the periods of largest discrepancy, and so their contribution to the MSE and MAE was again small. Third, the difference in behaviour among surgeons within a speciality was often not large. This seemed to be because the NSW waiting list reduction program in 1995 (Shiraev and McGarry 1996) brought everyone's waiting times down to similar levels. Large differences did emerge at the end towards the end of the three year period. However, for surgeons with long waiting times, these months of large discrepancy were often not included in the analysis because a high proportion of patients were still on the waiting list. One might expect larger differences in accuracy to emerge if a study analysed data collected over a period during which differences between surgeons had not been reduced.

The study suffers from several limitations. First, the three years analysed might be regarded as atypical because of the waiting list reduction (WLR) program that ran during 1995, and because the introduction of a new urgency category in July 1997 affected how many patients were assigned to urgency category U8. It is unclear in which direction the WLR program affected the results. It appeared to reduce the differences between statistics derived from surgeon and specialty data. But, because it introduced greater changes over time, which the functions did not closely track, the observed levels of accuracy may be worse than they might otherwise have been. Nonetheless, the conclusions of the study were based on aspects of the statistics' performance that were consistent across all surgeons regardless of the individual characteristics. The impact of the introduction of the new category was also limited because, for many surgeons, the interval over which the analysis was performed ended prior to its introduction.

The other principal weakness was the inclusion of patients who were still on the waiting list. Their eventual waiting times were underestimated, but as their waiting times were already long in comparison to most other patients, the probable effect of this would be to underestimate the forecast error. Thus, the results of the analysis would be conservative. Another potential bias from the inclusion of these patients was the assumption that all would be admitted. A review of the proportion of admitted and removed patients with waiting times comparable to the 'still waiting' patients was therefore conducted. For all but two surgeons, more of these patients were admitted than removed. Consequently, the bias that resulted from including patients who would eventually be removed was considered to be minimal.

Appendix: estimation of minimum distance

The first step in estimating the minimum significant difference between two estimates of expected waiting time was to estimate the standard deviation (SD) of the forecasts from the MA3(TH) and MA1(CS) functions. This was estimated from the differences between successive terms, using the following equation that strictly applies only when values are normally distributed (Bissell 1994):

$$\sigma = \sqrt{\left[\frac{1}{2(n-1)}\sum_{i=2}^{n}(x_i - x_{i-1})^2\right]}$$

where x_i is the ith statistic in the sequence (i=1..n) of average waiting time statistics.

The advantage of this estimator is that removes much of the influence of any 'medium term' trend or shifts in the local mean.

The next step was to check whether the SD estimates were related to the average level of the series. For the MA3(TH) series, there was a strong relationship across the surgeons between the value of the SD and the mean value of the series (Pearson's r = 0.85). Fitting a simple regression model, the linear relationship between the two factors was estimated as:

SD = 0.107 Avg + 3.473 $r^2 = 0.73$

For the MA1(CS) series, there was no relationship (linear or otherwise) between the values of the SD and the mean value of the series (Pearson's r = -0.1). Thus, again for simplicity, the SD was defined to be 13.46 days, the average of the SD values across all surgeons. Finally, the minimum distance was constructed using the standard formula for the confidence interval about the difference of two sample means drawn from different Normal distributions of known variance:

$$D = 2.Z_{1-a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where Z is the reliability co-efficient for a confidence level of 100(1-a/2) percent.

For the MA1(CS) series, the standard error component in the above formula reduced to the square root of 2 times the standard error estimate. For the MA3(TH) series, it was decided to ignore the constant in the regression equation and assume that the standard error was simply 10% of the series average. Consequently, the standard error component of the above equation reduced to the square root of 2 times the predicted standard error of the midpoint between the two averages. The value of Z was chosen to be 1.96.

References

Armstrong PW 2000, 'First steps in analysing NHS waiting times: avoiding the 'stationary and closed population' fallacy.' *Statistics in Medicine*, vol 19, pp 2037-2051.

Bissell D 1994, *Statistical methods for SPC and TQM*, Chapman and Hall, London.

Black N 1998, 'Potential biases were not taken into account in study of waiting times', *BMJ*, vol 316, pp 149.

Box G and Luceno A 1997, *Statistical control by monitoring and feedback control*, Wiley and Sons, New York.

Cromwell DA, Griffiths DA 2002, 'The implications of waiting list behaviour for interpreting waiting time statistics', *Australian Health Review*, vol 25, pp 40-49.

Cromwell DA, Griffiths DA, Kreis IA 2002, 'Surgery dot.com: the quality of information disseminated by web-based waiting time information services', *Medical Journal of Australia*, vol 177, pp 253-255.

DeGroot MH 1986, *Probability and Statistics*, Addison Wesley, Reading, Massachusetts.

Don B, Goldacre MJ, Lee A 1987, 'Waiting list statistics 3 : comparison of two measures of waiting times', *BMJ*, vol 295, pp 1247-1248.

Makridakis S, Wheelwright SC, and Hyndman RJ 1998, *Forecasting: methods and applications*, Wiley and Sons, New York.

Mason A 1976, 'An epidemiological approach to the monitoring of hospital waiting lists', *Proceedings of the Royal Society of Medicine*, vol 69, pp 939-942

Mordue A, Kirkup B 1989, 'An appraisal of waiting list problems', *Health Trends*, vol 21, pp 110-113.

Pugh EJ 1987, 'Visual presentation of waiting list statistics', *Hospital and Health Services Review*, vol 83, pp 111-114.

Shiraev N, McGarry J 1996, 'Waiting list reduction program: results to November 1995', *NSW Public Health Bulletin*, vol 6, pp 147-150.

White A 1980, 'Waiting lists: a step towards representation, clarification, and solving information problems', *Hospital and Health Services Review*, vol 76, pp 270-274.