

PERFORMANCE MONITORING OF MEDICAL OUTCOMES

Methodological Issues in Recent Developments

Vern Farewell

MRC Biostatistics Unit, Cambridge, UK

PERFORMANCE MONITORING

Broad Aims:

- Find out 'what works'
- Identify functional competence of practitioners or institutions
- Public accountability

Much of the thinking about performance monitoring has borrowed from industrial setting. It is interesting to note that in that setting there has been a shift from the use of PM as a means of detecting failure to its use in the positive role of encouragement of constant improvement by evolutionary changes.

WHAT CAN WE MONITOR?

Performance Indicators: Good, Bad and Ugly

RSS Working Party on Performance Monitoring in the Public Services

(Sheila Bird, David Cox, Vern Farewell, Harvey Goldstein, Tim Holt, Peter Smith)

Eleven recommendations made.

Selected Recommendations from RSS Working Party

2. A PM procedure must have **clearly specified objectives** and be **methodologically rigorous**. Individuals and institutions monitored should have *substantial* input to the development of the PM procedure.
3. A PM procedure should be designed to discourage counter-productive behaviour.
6. Performance Indicators need clear definition. They are subject to **variation due to systematic factors** (e.g. mix of students) **and simple randomness**. This must be recognised in their use.
7. The reporting of PM data should always include **measures of uncertainty**.

1. PM procedures need a protocol.
2. PM procedures need clear objectives and methodological rigour.
3. PM procedure should discourage counter-productive behaviour.
4. Cost-effectiveness should be considered.
5. Independent scrutiny of a PM procedure safeguards public accountability, methodological rigour, and individuals/institutions monitored.
6. PIs need clear definition. Variation, essential and systematic, must be recognized.
7. Reporting of PM data should include measures of uncertainty.
8. Aspects of PM should be investigated under research council sponsorship.
9. Research needed to provide robust methods to evaluate Government policies/initiatives. (Randomized trials?)
10. Ethics should be addressed.
11. Education is needed about role and interpretation of PM data.

Can You Make Sense of This?

By grade national comparison										
Subject	number sa	A*	A	B	C	D	E	F	G	U
Art + design (APP)	7854	5	9.8	17.7	22.6	17.4	11.2	7.7	5.6	3
Art	212357	6.8	15.3	21.1	28.3	13.4	8.5	4.5	1.8	0.3
3510-Art & Design	58	8.6	15.5	20.7	32.8	3.4	5.2	10.3	3.4	0
Biology	60082	18.1	26.1	25.2	19	7.6	2.2	0.8	0.3	0.7
1010-Biology	22	63.6	18.2	18.2	0	0	0	0	0	0
Chemistry	56764	20.1	25.6	24.9	19.6	6.8	1.9	0.4	0.2	0.5
1110-Chemistry	22	59.1	22.7	18.2	0	0	0	0	0	0
DT (all)	371672	4	12.9	16.1	25.6	19.3	10.4	5.7	3	3
9020-D&T Food Tech	65	9.2	23.1	21.5	29.2	10.8	4.6	0	1.5	0
9030-D&T Graphic Pr	30	3.3	10	20	20	13.3	16.7	10	3.3	3.3
9040-D&T Resistant M	32	0	0	15.6	34.4	28.1	3.1	12.5	6.3	0
9050-D&T Textiles Te	42	16.7	26.2	11.9	23.8	14.3	7.1	0	0	0
Drama	100808	5.5	17.1	26	22.8	14.2	7.8	3.9	1.9	0.8
5210-Drama	44	0	6.8	45.5	38.6	6.8	0	0	2.3	0
Eng lit	572161	4.8	14.8	23.1	25.1	15.7	8.9	4.1	1.7	1.8
5110-English Literat	180	6.7	15	30.6	23.9	12.8	7.2	0.6	0.6	2.8
English	721762	3.9	11.3	19.4	27	20	10.2	4.8	2	1.4
5010-English Langua	159	1.9	13.8	29.6	30.8	7.5	7.5	2.5	2.5	3.8
French	236189	9.6	13.2	17.7	24.2	17.5	9.7	5.1	2.4	0.6
5650-French	58	17.2	13.8	10.3	29.3	8.6	13.8	3.4	1.7	1.7
Geography	213469	9.2	14.9	17.4	24.4	15.3	8.7	5	2.7	2.4
3910-Geography	68	7.4	14.7	17.6	29.4	22.1	1.5	4.4	2.9	0
German	90311	8.6	13.9	18.8	28	16.7	7.7	4	1.9	0.4
5670-German	72	8.3	19.4	16.7	23.6	9.7	12.5	6.9	2.8	0
Health (APP)	30478	0.6	5.6	15.7	23.3	18.6	14.4	10.6	6.9	4.4
0003-Health & Social	22	0	0	0	27.3	27.3	0	36.4	9.1	0
History	231657	10.3	18.2	20.1	18	12.8	9.1	6	3.3	2.2
4010-History	110	5.5	23.6	25.5	25.5	10.9	4.5	0.9	2.7	0.9
Home Economics	46528	3.6	9.1	13.6	27.5	19.1	12.7	7.6	3.8	3
3310-Home Economic	7	0	0	0	57.1	28.6	14.3	0	0	0
IT (APP)	44554	0.6	5.4	16	22.6	16.7	13.3	11	8.2	6.2
0010-Information Tec	112	0	8.9	10.7	26.8	25	17.9	8.9	1.8	0
Latin (classical subj	16305	30.8	26.8	17.2	13.1	6.7	2.3	1	0.6	1.5
6610-Latin	11	54.5	27.3	0	9.1	0	0	0	0	9.1
Leisure and Tourism	18142	0.2	2.5	8.7	18	19.2	16.6	14.4	11.3	9.1
0004-Leisure & Touris	20	0	0	10	10	30	10	30	10	0
Mathematics	750570	4.2	9	17.6	23.5	18.1	13.3	7.3	3.2	3.8
2210-Mathematics	194	6.2	15.5	20.1	21.1	13.4	11.9	7.7	3.1	1
Music	60668	9.6	20.3	24	19.5	10.8	7	4.1	2.4	2.3
7010-Music	28	10.7	25	32.1	10.7	3.6	3.6	3.6	10.7	0
Physics	56035	20.7	26.1	24.1	19.7	6.9	1.6	0.4	0.1	0.4
1210-Physics	22	50	22.7	27.3	0	0	0	0	0	0
RE	159681	11.2	19.3	21.3	18.8	12.2	7.8	4.7	2.8	1.9
4610-Religious Studie	20	5	40	30	10	10	5	0	0	0
Sci Voc (APP)	27471	0.1	1.4	7.2	25.2	27.4	20	11.3	4.9	2.5
0008-Science (Voc)	48	0	16.7	8.3	0	12.5	16.7	33.3	12.5	0
Sci Double	959578	5.4	9.2	14.7	29	19.4	12.1	6.8	2.9	0.5
1370-Science Double	252	11.1	17.5	19	18.3	19	9.5	2.4	2.4	0.8
Sci Single	96374	0.4	1.8	4.3	18.2	20.8	21.2	17.8	9.2	6.3
1310-Science Single	19	0	0	0	5.3	0	15.8	31.6	36.8	10.5
PE (sport)	152826	6.1	14	20.3	20.9	23.8	10.3	3.3	1	0.3
7210-Sport/PE Studie	38	2.6	10.5	31.6	23.7	31.6	0	0	0	0
All Subjects GCSE	5752152	6.3	12.8	18.3	25	17.3	10.2	5.6	2.6	1.9

FUNNEL PLOTS (Binomial)

- r_i A* to C grades from n_i students taking subject i
- I subjects
- $p_i = r_i/n_i$ and $\bar{p} = \Sigma p_i / \Sigma n_i$
- $\sigma^* = \sqrt{\bar{p}(1 - \bar{p})}$ and $z_i = (p_i - \bar{p}) * \sqrt{n_i} / \sigma^*$
- Allow for overdispersion: $\phi = \Sigma z_i^2 / I$ with "Trimming or Winsorization"
- Plot limits (say 95% and 99%) based on standard errors $\sqrt{\phi} \sigma^* / \sqrt{n}$

School Results (A*-C)

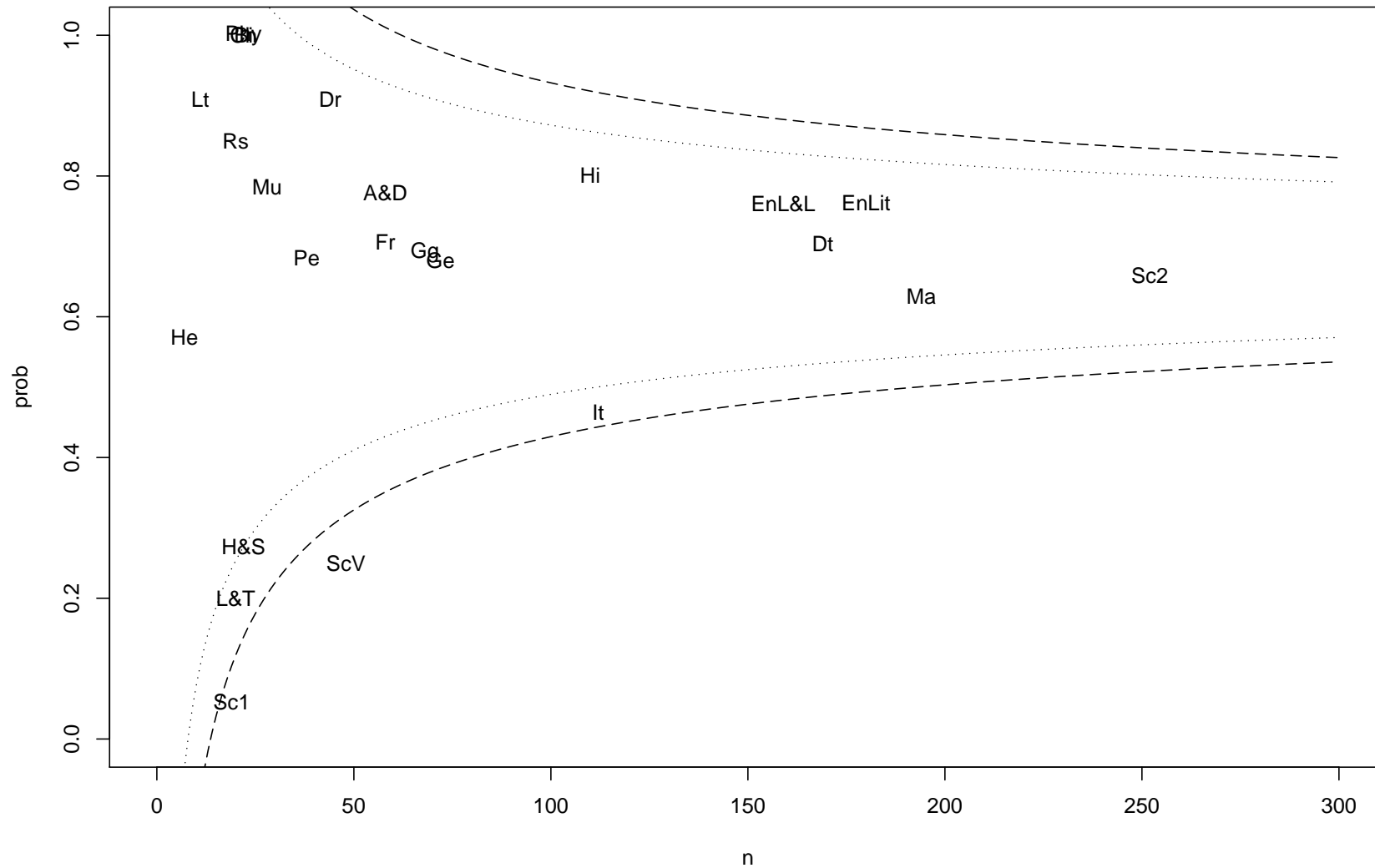


Figure 2: Internal School Results

FUNNEL PLOTS (Poisson)

- National A* to C proportion in subject $i = p_i^{\text{national}}$
- Expected in school: $e_i = n_i p_i^{\text{national}}$
- Plot Observed/Expected (r_i/e_i) vs Expected (e_i)
- Use Poisson limits

Results vs National (A*-C)

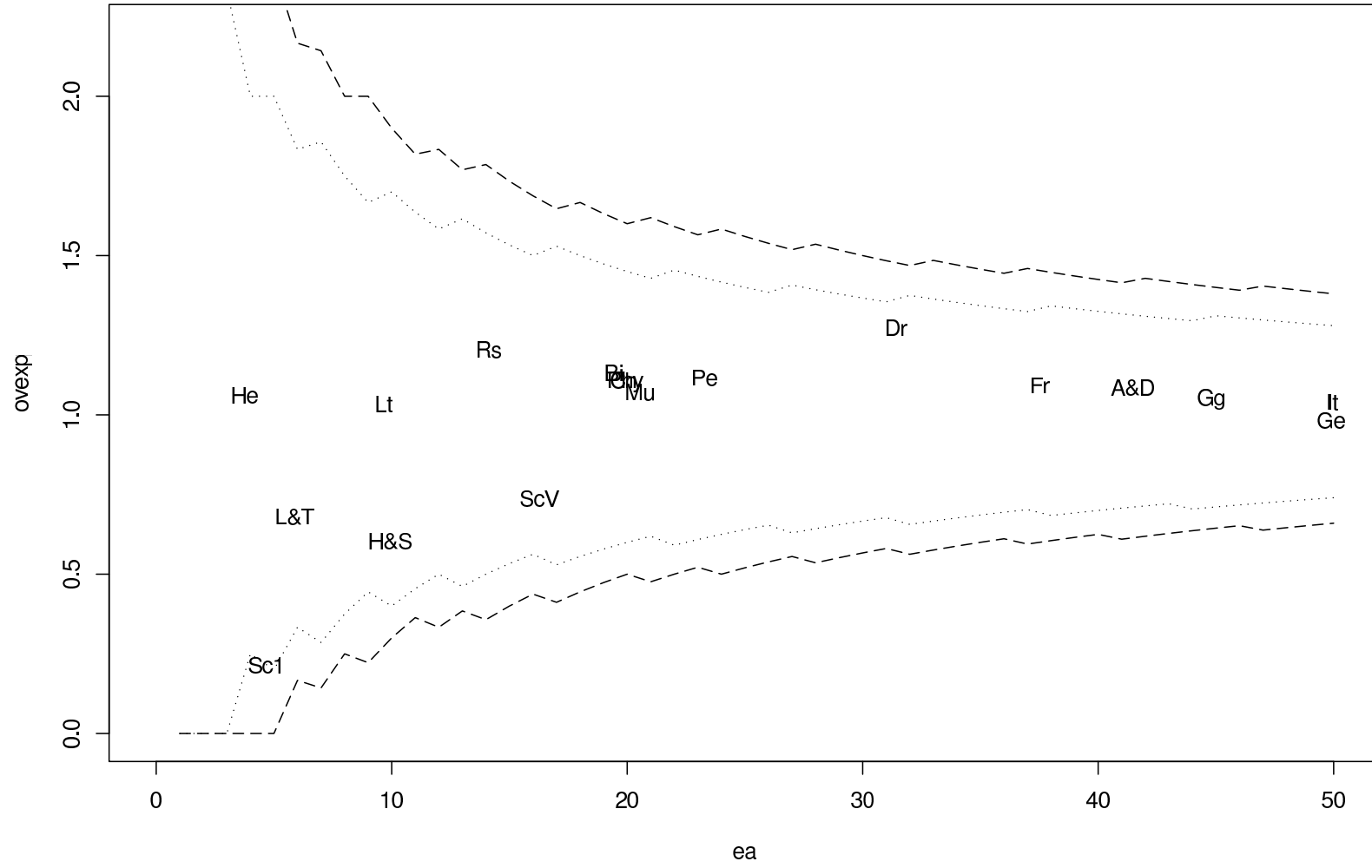


Figure 3: School vs National Results

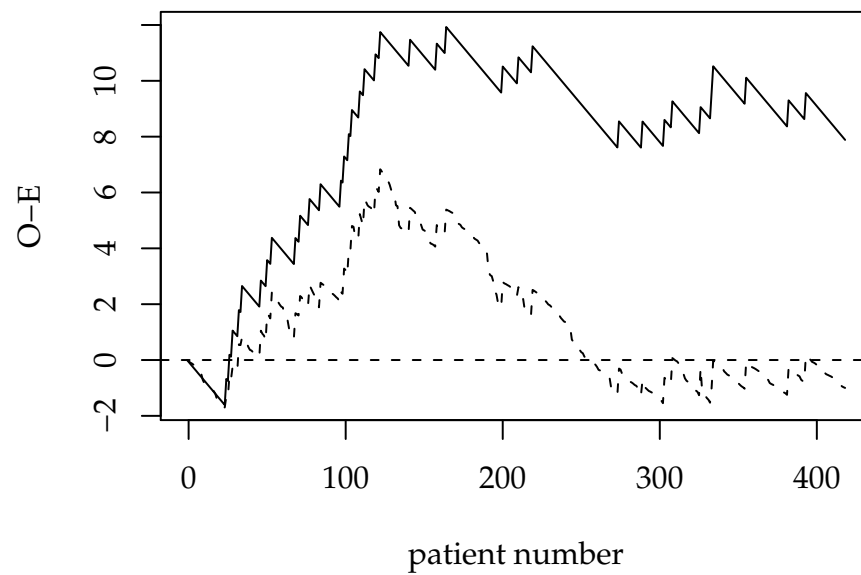
Risk Adjusted Longitudinal Monitoring

- Great intuitive appeal to a graphical display of *observed* performance to *expected* performance.
- The need for risk adjustment (case-mix adjustment) is now well accepted.
- Initially, for monitoring of surgical mortality (binary outcome), Poloniecki *et al* and Lovegrove *et al* have termed **risk adjusted O-E plots** as Cumulative Risk Adjusted Mortality (CRAM) charts and Variable Life Adjusted Display (VLAD) respectively.
- O-E plots can also be used for non-binary data

EXAMPLE 1

Cardiac Surgical Mortality:

- Data for a single surgeon over four years.
- Binary outcome for t^{th} patient, Y_t , is 0-1 indicator of death within 30 days of surgery.
- For each patient, there is an estimated probability of surgical mortality, p_t .
- $O - E$ plot is plot of $\sum_1^T (Y_i - p_i)$ versus T .



h

Figure 4: O-E plot of cardiac surgical outcomes (— unadjusted; - - risk-adjusted)

EXAMPLE 2

Deaths in the GP practice of Harold Shipman

- Use deaths in six month periods. Assumed to follow Poisson distribution with mean λ .
- Risk adjustment through calculation of age and sex adjusted rate for the patient mix in Shipman's practice, averaged over the years to give a single rate used throughout the chart.
- Expected rate is calculated as 35 deaths/year for all patients and 12 deaths/year for females over 75 years old.

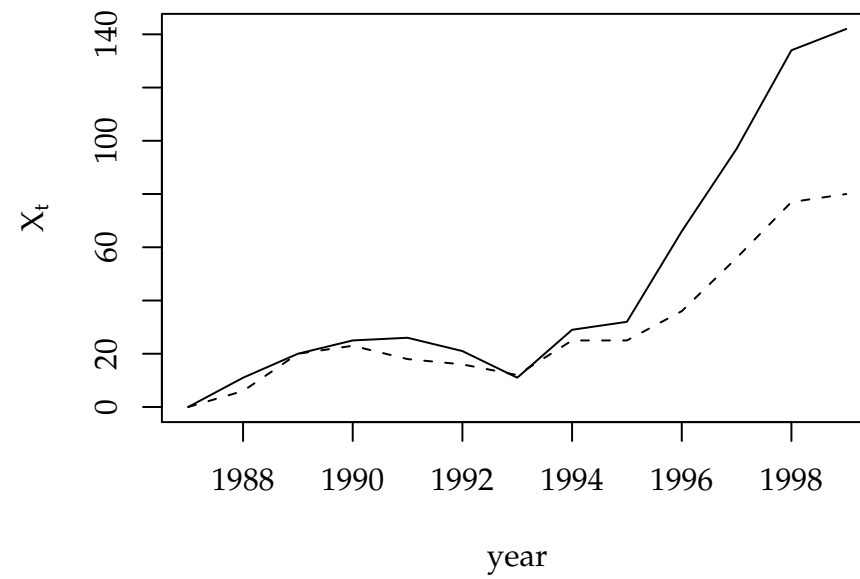


Figure 5: O-E plot of deaths under Harold Shipman (1987-1998), where the expected value is based upon the England and Wales average (—all patients; - - females aged 75 and over)

BUILDING ON THE PAST

- Classical SPRT:

- Designed to test hypothesis H_0 vs H_1
- Plots

$$X_t = X_{t-1} + W_t, \quad t = 1, 2, 3, \dots$$

at t^{th} observation. [$X_0=0$]

–

$$W_t = \log\left(\frac{L_{1t}}{L_{0t}}\right)$$

where L_{0t} is likelihood contribution for t^{th} observation and is proportional to the probability of the outcome under H_0 . L_{1t} is defined similarly.

- For Poisson, binomial and Bernoulli data, W_t can be written in the form $\mu O - \nu E$.
- No natural model for binomial, Bernoulli or Poisson data leads to O-E weights
- SPRT has two absorbing barriers

$$a = \log\left(\frac{\beta^*}{(1 - \alpha^*)}\right)$$

$$b = \log\left(\frac{(1 - \beta^*)}{\alpha^*}\right)$$

where α^* and β^* correspond to upper limits for the type I and type II errors of the test H_0 vs H_1 .

Log-likelihood Weights

	<i>No. of failures $Y_t \sim$</i>		
	Poisson(λ)	Binomial(n, p)	Bernoulli(p)
<i>Measure of difference in performance R_1</i>	RR = λ_1/λ_0	OR = p_1q_0/p_0q_1	OR = p_1q_0/p_0q_1
<i>Weights W_t</i>	$Y_t \log(R_1) - \lambda_0(R_1 - 1)$	$Y_t \log(R_1) - n \log(1 - p_0 + R_1 p_0)$	$Y_t \log(R_1) - \log(1 - p_0 + R_1 p_0)$
<i>Expected value E</i>	λ_0	np_0	p_0
<i>W_t in terms of $O-E$</i>	$\mu O - \nu E$	$\mu O - \nu(p_0)E$	$\mu O - \nu(p_0)E$

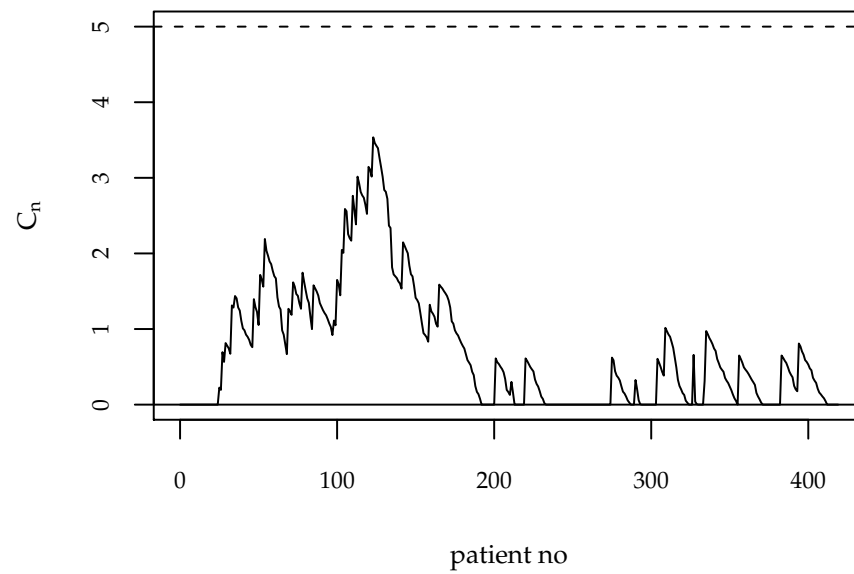
Table 1: Form of log-likelihood ratio weights for charts detecting changes in risk for Poisson, Binomial and Bernoulli data types. Observed value, O , is Y_t . For all three data types, $\mu = \log(R_1)$. For Poisson, $\nu = R_1 - 1$. For Binomial and Bernoulli, $\nu(p_0) = \log(1 - p_0 + R_1 p_0)/p_0 \approx R_1 - 1$

Tabular CUSUM

- Involves plotting

$$X_t = \max(0, X_{t-1} + W_t) \quad t = 1, 2, 3, \dots$$

- Holding barrier at $X_t = 0$, rather than a lower absorbing barrier
- CUSUM is said to signal when $X_t > h$
- Not a test of H_0 vs H_1
- Assumes H_0 is true at beginning of monitoring and looks for evidence of a change
- Logic is sensible for routine monitoring for which termination of a chart in favour of the null hypothesis makes little sense
- Optimal choice of weight W_t is based on log-likelihood ratio. (Moustakides, 1986)



h

Figure 6: Risk-adjusted CUSUM of cardiac surgical outcomes, $h = 4.5$, in control ARL=6700

OTHER POSSIBILITIES

- *Resetting SPRT (RSPRT) Chart*
 - Formally propose a sequence of SPRTs with an absorbing barrier at b and a ‘resetting’ barrier at a . Values of α^* and β^* are used to set boundaries but have no relation to error rates
 - CUSUM with likelihood ratio weights is equivalent to a RSPRT chart with barriers close to $(0, h)$
 - Both CUSUM and RSPRT have formal type I and type II error rates of $\alpha = 1$ and $\beta = 0$ reflecting the fact that the upper boundary will eventually be crossed

OTHER POSSIBILITIES

- *The Shewhart chart*
 - Simply charts actual observations (sometimes standardized) of a process
 - Process deemed ‘out of control’ when prespecified probability limits are crossed (Usually 99% or 3σ limits set)
 - Usually concern is only with large changes in the process
 - To provide a sensible chart, binary outcomes would need to be grouped to give binomial data and what is known as a Shewhart p -chart. Limits often based on normal approximations
 - To allow for risk-adjustment, the probability of failures are allowed to vary by case. A simple adjusted chart might then assume binomial data with an average value for p over any particular time period.

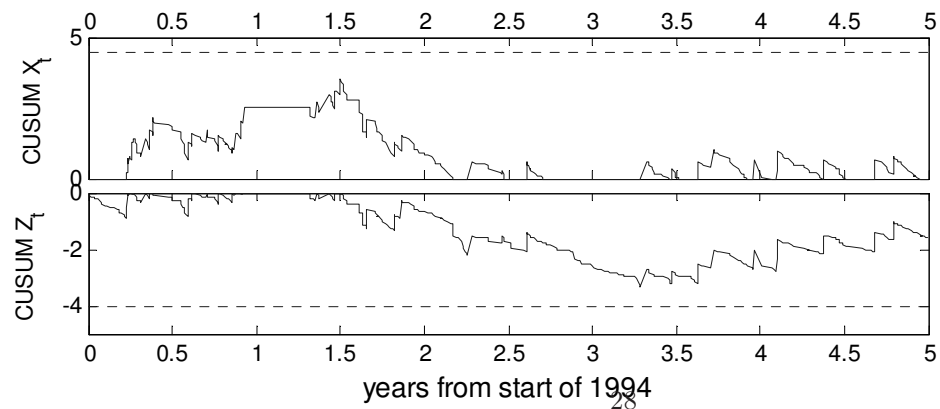
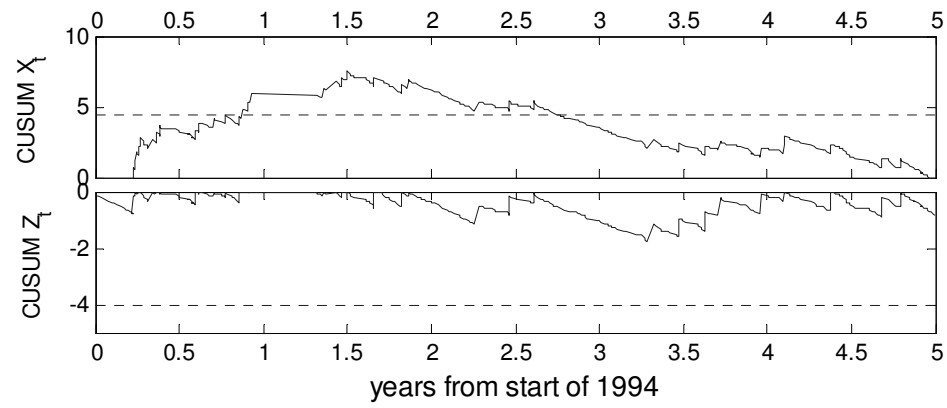
OTHER POSSIBILITIES

- *The SETS method*
 - First introduced by Chen for surveillance of congenital malformations (a binary outcome).
 - Method is based on the number or ‘set’, X , of observations after a failure up to and including the next failure. If this set is $\leq T$ on n successive occasions, an alarm is signalled.

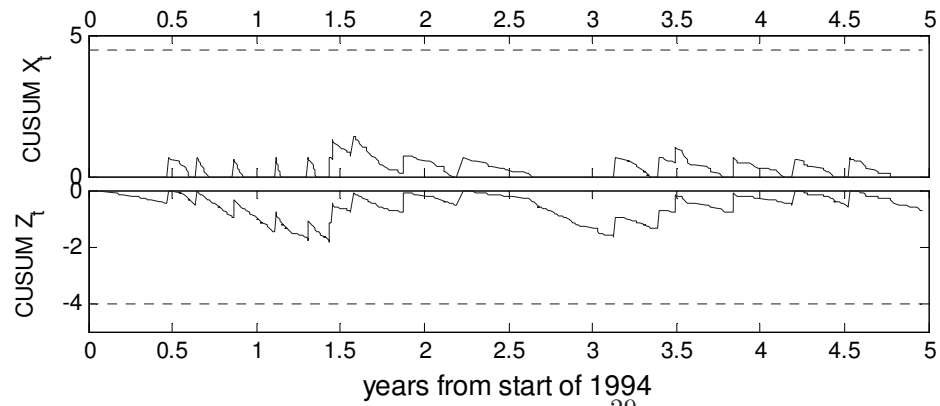
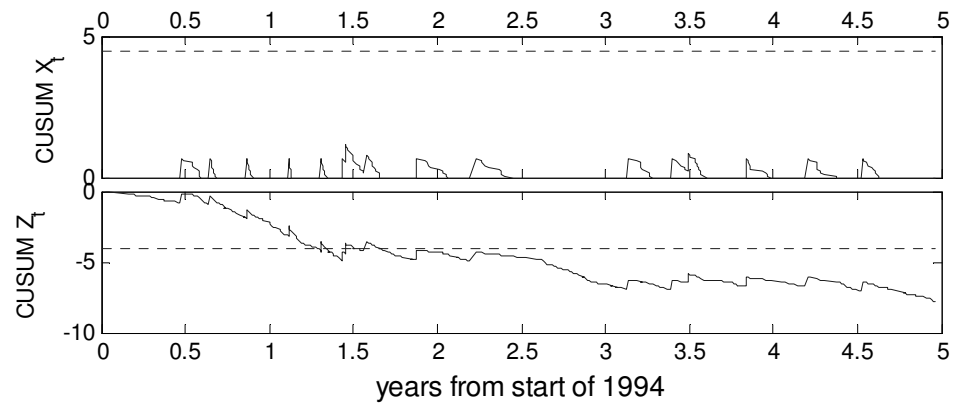
FOR WHAT ARE WE LOOKING?

- Primary motivation is usually to detect a deterioration in performance.
- There is some standard against which this is measured.
- Also important to be able to detect improvements, especially with long term monitoring. This may prompt reevaluation of the standard (which perhaps should be done on a regular basis in any event) or identify centres or individuals who do particularly well and may have transferable methods of working.
- Many of the monitoring techniques can be used to detect either deterioration or improvement. Therefore, it is often useful to set up a monitoring process for each.

Experienced surgeon



Trainee surgeon



After a problem has been detected and dealt with:

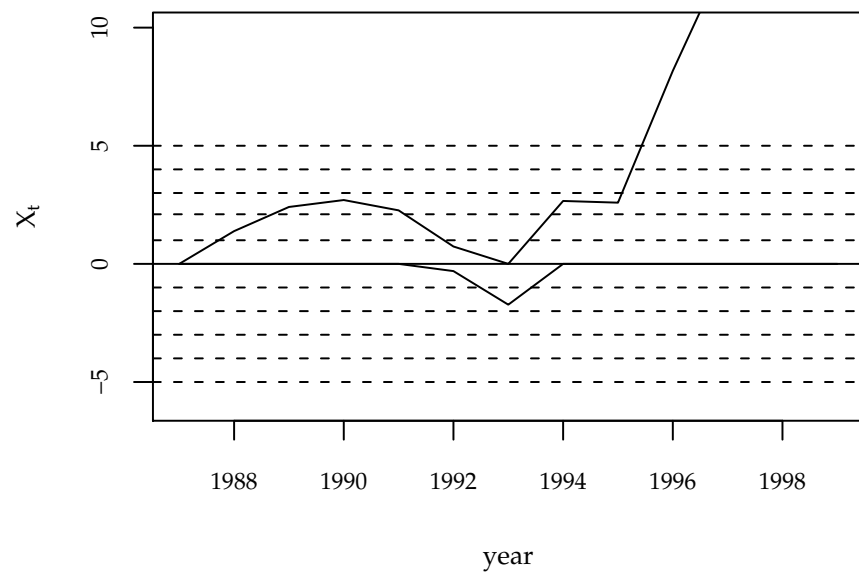
- Usually not sure that the problem is solved.
- Perhaps not reasonable to assume H_0 when restarting monitoring.
- Use techniques like Fast Initial Response CUSUMs which are CUSUMS which start a point in between $(0, h)$ rather than at zero. ($h/2$ is a common choice.)
- Equivalent to a RSPRT with the first SPRT in the sequence having boundaries at $(-r, h - r)$ and the rest having $(0, h)$.
- The question the first SPRT answers is that of a classical SPRT. It chooses between H_0 and H_1 .
- This is a particular example of the need to choose a chart based on context as well as any statistical properties it might have.

COMPARING METHODS

- Has to depend on context and expected use of the chart.
- However, a useful starting point is often average run lengths (ARLs).
- ARL under H_0 is analagous to a type I error rate. ARL under H_1 is analagous to power. Thus ARLs often used to set boundaries.
- ARLs sometimes calculated analytically, sometimes via (discrete) Markov Chain approach, sometimes by simulation.
- ARLs not always sufficient. May want to look at distribution of run lengths.
- In control run lengths are usually approximately geometric but there is no comparable general results for out of control run lengths.

Shipman Data: All patients.

- Use two-sided CUSUM. Assume 20% change in rate is important.
- A boundary of $h = 3$ would generate 1 false-positive or false-negative signal over 52 years
- Chart signals in 1995 for any boundary in range $[3,8]$
- Use of all data leads to $\hat{\lambda} = 42.3$ which is bias adjusted to 41 [CI:(37,45)] corresponding to a 17% increase in rate.
- Use of data since 1992 (chart last at zero then) leads to $\hat{\lambda} = 53.3$ which is bias adjusted to 52 [CI:(43,60)] corresponding to a 49% increase.



h

Figure 9: CUSUM monitoring death rates per year under Harold Shipman, 1987-1998

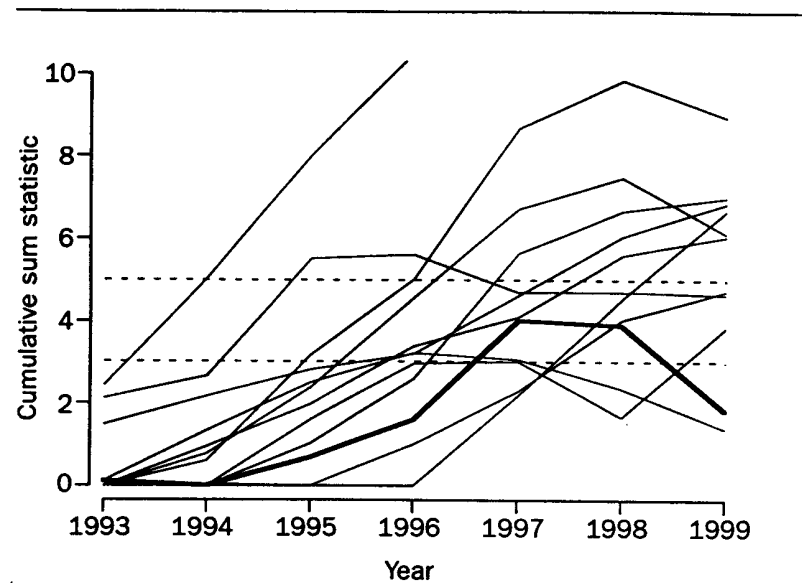


Figure 1: Cumulative sum charts for 12 family physicians signalling at any time between 1993 and 1999
 Charts designed to detect 4 SD increase in standardised excess mortality ($K=4-0$) with an alarm threshold of $h=3$ ($h=5$ is also shown). Bold line=Harold Shipman's cumulative sum chart.

Figure 10: Best et al

MULTIPLE PLOTS

Multiple Outcomes:

Approaches to this can be developed. *e.g.* Simultaneous CUSUM (SCUSUM) charts of Steiner *et al*

MULTIPLE PLOTS

Multiple Processes:

- Need to think through the relevance of standard approaches to multiplicity. E.G. How often will the protection of an ‘experimentwise error rate’ provided by a Bonferroni type procedure be relevant?
- Approach should depend critically on what happens when a ‘signal’ occurs.
- Some compromise between protection of medical practitioners and patients seems inevitable.
- ARLs might help to characterize the effect of multiplicity
- The use of the concept of False Discovery Rates also warrants exploration.

Recent Work

OA Grigg and DJ Spiegelhalter

- Considering null steady-state distribution of CUSUMs.
 - Allows calculation of a p-value based on the proportion of time a CUSUM will be above a specified value.
 - Can then make use of FDR methodology.
- Estimation of the “level” of a dynamic risk-adjusted process.
 - Have developed risk-adjusted exponentially weighted moving average methodology.