Qualities of risk tools; the evidence base (for PEWS discussions)

A very brief briefing to provide shared understanding of terminology and challenges

A PEWS is intended to do a better job than unsupported clinical judgement in identifying babies and children whose condition may be deteriorating, and therefore need senior review. The challenge of using scores or thresholds to help identify patients at risk is common to many areas of healthcare (e.g. risk of stroke after TIA, risk of falling in hospital, likelihood of injured ankle being a fracture). There are some essential qualities of all such tools that are important to consider in discussions on PEWS.

What is the best way of developing an accurate tool?

The best way of developing such tools is first undertaking logistic regression (for PEWS, taking samples of children whose condition deteriorated and comparing to children who did not, to find what characteristics significantly differ and should be included in the PEWS) followed by prospective validation (when used in clinical practice, how well does the tool identify children whose condition is deteriorating).

Sensitivity and specificity matters (Figure 1)

Only if the tool is tested for predictive validity can we know if it does a better (or worse) job of triggering appropriate review of deteriorating children than clinical judgement alone. It is extremely easy to develop a tool that is highly sensitive (an extreme example of a highly sensitive PEWS would be *'unless the child is out of bed and playing, they could be deteriorating'*) OR is highly specific (an extreme example of a highly specific PEWS would be *'children who have recently become unconscious are deteriorating'*). The best tools have high sensitivity AND high specificity (not too many deteriorating children missed, and not too many 'false alarms'). Statistical modelling is normally required to find the threshold that gets the best balance.

There are examples in the literature where unsupported clinical judgement has been found to be equally or more accurate than a complex tool, or where a very simple rule has better predictive qualities than complex rules derived over decades of research (e.g. *'are they unsteady on their feet?*' performs better than STRATIFY for risk of falls).

Other important qualities

Any risk tool needs to be easy to complete, with a design that minimises human error, and have perceived clinical value (i.e. in pilots, frontline staff believe the gain from completing it is worth the effort). It also needs good inter-rater reliability (i.e. in context of PEWS, two people using it on the same child would reach the same threshold).

But isn't any tool better than no tool?

If the sole purpose of a tool is to raise awareness of an issue, there are potentially other ways of building this into standard practice (e.g. as falls risk scores were taken out of use, they were replaced with an admission question of *'Is the patient or their family worried they might fall?'*) Parallels in the context of PEWS might be incorporating into standard observation charts a prompt for questions such as *'Is the child recovering as quickly as expected? Are their parents becoming more worried? Do they need urgent medical review?'*

Manage the risks of new tools

Whilst all such risk tools are intended as an aid to clinical judgement, there is usually an unintended effect of it overriding clinical judgement – staff who were concerned about a child reassuring themselves their concern was unwarranted because the score is low – that needs education and management as the tool is introduced.

Quality improvement techniques used to embed use of any risk tool also need to be mindful that in measuring completion of the tool, there is a risk that completing it becomes seen as an end, not a means to an end.

Figure One: terminology of predictive scores

	Actually deteriorating	Not actually deteriorating
Triggered on PEWS	**A**	**B**
Below threshold on PEWS	**C**	**D**

Green areas are where the PEWS is correctly identifying level of risk

Red represents missed opportunities to spot a deteriorating child

Amber represents 'false alarms'

Sensitivity (true positive) = A/A+C

"What percentage of children who WERE deteriorating 'triggered' on PEWS?

Specificity (true negative) = D/B+D

"What percentage of children who WERE NOT deteriorating were 'below threshold' on PEWS?"

Positive Predictive Value (PPV) = A/A+B

"Of all the children who 'triggered' on PEWS, what percentage WERE deteriorating"

Negative Predictive Value (NPV) = D/C+D

"Of all the children below threshold on PEWS, what percentage WERE NOT deteriorating?"

Frances Healey, RGN, RMN, PhD,

Senior Head of Patient Safety Intelligence, Research & Evaluation,

NHS England

08 December 2014