Facility versus unit level reporting of quality indicators in nursing homes when performance monitoring is the goal

Peter G Norton, Michael Murray, Malcolm B Doupe, Greta G Cummings, Jeff W Poss, Janet E Squires, Gary F Teare, Carole A Estabrooks

ABSTRACT

Objectives: To demonstrate the benefit of defining operational management units in nursing homes and computing quality indicators on these units as well as on the whole facility.

Design: Calculation of adjusted Resident Assessment Instrument – Minimum Data Set 2.0 (RAI–MDS 2.0) quality indicators for: PRU05 (prevalence of residents with a stage 2–4 pressure ulcer), PAIOX (prevalence of residents with pain) and DRG01 (prevalence of residents receiving an antipsychotic with no diagnosis of psychosis), for quarterly assessments between 2007 and 2011 at unit and facility levels. Comparisons of these risk-adjusted quality indicators using statistical process control (control charts).

Setting: A representative sample of 30 urban nursing homes in the three Canadian Prairie Provinces.

Measurements: Explicit decision rules were developed and tested to determine whether the control charts demonstrated improving, worsening, unchanging or unclassifiable trends over the time period. Unit and facility performance were compared.

Results: In 48.9% of the units studied, unit control chart performance indicated different changes in quality over the reporting period than did the facility chart. Examples are provided to illustrate that these differences lead to quite different quality interventions.

Conclusions: Our results demonstrate the necessity of considering facility-level and unit-level measurement when calculating quality indicators derived from the RAI–MDS 2.0 data, and quite probably from any RAI measures.

INTRODUCTION

In 2002, Nelson et al. began to document the importance of clinical Microsystems. Informed by the organisational literature, they defined these Microsystems as “small, functional, front-line units that provide most healthcare to most people” and “the place where patients and providers meet. The quality and value of care produced by a large health system can be no better than the services generated by the small systems of which it is composed” (ref. 1 p 473). Quality improvement efforts in Great Britain, Sweden, Australia and Canada have focused on clinical Microsystems with promising early evaluations. For example, a leading children’s hospital in the USA has embraced this concept by creating ‘microsystem leaders,’ accountable for quality and safety in their own Microsystems.

In this article, we demonstrate that measurement at the microsystem level in nursing homes is feasible, enhances opportunities for quality assurance and improvement and is scientifically sound. The article enhances our previous work that gave a practical and easily applied method for identification of clinical Microsystems in nursing homes. The results should be useful to those managing and working in the sector as well as to researchers.

Translating Research in Elder Care (TREC) is an ongoing programme of research focused on improving the quality and safety of care delivered to residents of nursing homes. Protocols for the present programme phase have been published elsewhere. For over 6 years, TREC has closely
followed a representative cohort of 30 urban nursing homes in the Canadian Prairie Provinces. These homes range in size from 48 to 446 beds (mean of 132) and have a representative distribution of owner-operator models with eight being publically owned and operated, 14 being run by the voluntary sector and eight being private for profit. TREC has collected data on organisational context in these homes on two occasions\(^8\) and has captured Resident Assessment Instrument – Minimum Data Set 2.0 (RAI–MDS 2.0) data from the nursing homes from 2007 onward. The present report focuses on this latter data and their use for our purposes of quality assessment and improvement.

As part of this work, we developed, validated and reported a functional definition of care unit in these homes\(^9\): a geographical area in a facility, serving a population of patients while they reside there, with dedicated management, which is characterised by:

- A regular group of care providers (eg, healthcare aides, licensed practical nurses, registered nurses) who deliver the direct care and who work most of their shifts (typically at least 60%) on one unit.
- A care manager who is in charge of the whole unit, but whose supervision may stretch across several units, for example, registered nurses on night shift.
- A nurse who oversees the unit on a shift-by-shift basis, but whose supervision may stretch across several units, for example, registered nurses on night shift.

As part of our validation, we demonstrated that critical elements of organisational context, as reported by frontline workers in the facility, best aggregate to these defined units.\(^9\) Furthermore, when we applied our definition to our sample,\(^9\) we found that in 28% of nursing homes, a realignment of the unit structures defined by facility management was needed to ‘fit’ our definition. One common reason for realignment was that ‘unit’ had been used to define ‘houses’—living groups—rather than actual management structures. We then checked our realigned unit definitions with facility management; in all cases, they confirmed that we had defined the units correctly from a management point of view. Finally, we mapped our definition of unit onto the definition of clinical microsystems given by Nelson \textit{et al}\(^8\) and achieved a very good fit. We concluded that the units we had defined were indeed the clinical microsystems of nursing homes. Given the recommendations that flow from present microsystem theory and our understanding of quality improvement, these units should be the focus of quality and safety improvement activity in homes.

In this article, we address the question: Can and should quality of care be assessed at the unit level in nursing homes?

Measurement is an essential component of quality improvement and has several key uses in that work\(^13\)\(^ 14\): identifying areas for improvement, evaluating the success of improvement efforts, assessing sustainability and scale-up of successful local improvements and reporting for accountability. All homes in our sample and the majority of nursing homes in Canada collect and report data quarterly using the RAI–MDS 2.0 system.\(^15\) This system has been studied for many years and quality indicators have been developed and validated.\(^16\)–\(^18\) More recently, work was performed to risk adjust many of those indicators to account for the differences in resident populations across facilities.\(^19\) In Canada, RAI–MDS 2.0 is the version used in residential long-term care settings, and the Canadian Institute of Health Information has endorsed 35 of these indicators for reporting quality data back to nursing homes.\(^20\)

In Canada, as in other jurisdictions, RAI–MDS 2.0 measures are reported back to nursing homes and others (eg, governments, funders, etc), with indicator data aggregated to the facility level without distinction by unit within a home. This type of aggregating blends results from multiple units such that detail necessary to manage a clinical microsystem can be lost. Unit-level improvement efforts need unit-level measurement.

In this article, we continue our exploration of clinical microsystems (units) in our representative set of nursing homes in the Prairie Provinces in Canada, with a focus on the role of microsystems in performance measurement. We examine several of the Canadian Institute for Health Information (CIHI) endorsed quality indicators, and demonstrate that homes could benefit substantially in their quality work by computing indicators at the facility and unit levels. We demonstrate:

- The value in computing RAI–MDS 2.0 indicators at the unit and facility levels (illustrated with examples);
- The proportion of homes that benefit from having RAI–MDS 2.0 indicators at both levels.

\section*{Methods}

In quality improvement work, temporal data such as quarterly RAI–MDS 2.0 indicators are often considered and evaluated using statistical process control (SPC) methods. A major tool for SPC is the control chart, containing two parts: (1) a series of measurements plotted in time order and (2) the control chart template—three horizontal lines called the centre line (typically, the mean), the upper control limit (UCL) and the lower control limit (LCL).\(^21\) Readers unfamiliar with SPC can consult several excellent sources on the science and interpretation of control charts.\(^22\)–\(^25\) We computed control charts for three adjusted RAI–MDS 2.0 indicators in a stratified random sample of our study population of nursing homes and their units.

We computed RAI–MDS 2.0 quality indicator values using published methods\(^19\) and the standard model parameters employed by the Canadian Institute of Health Information.\(^26\)–\(^28\) We carried out a covariate adjustment of the indicator in each unit or facility to account for the difference of its population from the standard model.
To compare quality indicators across units and facilities, we needed rigorous a priori techniques to determine whether a unit or facility improved over time, stayed the same or got worse. The Manitoba Centre for Health Policy produced the only report found in the literature on such a technique, in 2008.28 Informed by that work we developed an initial set of empirical rules to make decisions about temporal control charts. We then generated several sets of control charts for quality indicators, each of which comprised 20–25 charts. Using the original decision rules and the initial rules, three reviewers independently ranked the charts. After scoring, reviewers held a consensus conference to resolve differences and rules were modified as necessary.

Next, with a new set of charts and the modified rules, the process was repeated. We completed two further iterations of this process to arrive at the final rules which are provided in the online supplementary material 2).

Setting and sample
Thirty urban nursing home sites in the TREC cohort have a total of 94 units. The distribution is shown in table 1.

Since we were interested in comparison of unit and facility data, we focused on the 25 homes with two or more units. It seemed conceivable that homes with different numbers of units could perform differently, concerning variation in unit and facility behaviours. Hence we stratified the 25 homes into homes with two, three or four or more units and randomly selected two homes from each of these strata.

Measures
In previous work, we involved researchers and decision-makers to determine a set of practice-sensitive RAI–MDS 2.0 indicators.29 In this work, we examined control charts for three indicators: PRU05 (prevalence of residents with a stage 2–4 pressure ulcer), PAI0X (prevalence of residents with pain) and DRG01 (prevalence of residents receiving an antipsychotic with no diagnosis of psychosis). All indicators were risk adjusted. Using SPSS (20.0) we computed control charts for each indicator on each sampled facility and each unit in those facilities.30 One facility (17) was not included in the DRG01 analysis since it had reports from only two time periods. In all other cases, we had 13–18 observations from mid-2007 to the end of 2011. Control charts were graded with the decision rules to decide whether and what kind of change had occurred over the 13–18 quarters (3.25–4.5 years). Change classification was carried out by two of the experienced raters.

Ethics
Ethics and operational approvals were obtained from all participating investigators’ universities and from the participating sites, respectively.

RESULTS
Inter-rater agreement in the development of the decision rules
Table 2 illustrates agreement statistics after each iteration in the development of the decision rules. At the end of the fourth iteration, the rules had stabilised and we declared the rules as the final ones.

We tested these rules further on another series of control charts, using two experienced raters from the previous team and adding two naïve raters. We began with two training sessions, each involving 20 charts, and discussed differences. Then we carried out a formal test on a new set of 20 charts. We achieved an average pairwise agreement of 92% with a Krippendorff’s α of 0.804.

The sample of homes
The sample included smaller and larger homes with a variety of owner-operator models (table 3). Facility identifier numbers are anonymised and are called 3, 6, 13, 17, 26 and 28 in this report.

The control charts
Table 4 describes the classifications by the two experienced reviewers of the 45 control charts considered and
the final agreed-upon classification of control charts for each facility and its units.

Additional tables S1–3 (online supplementary material) demonstrate the results by indicator. Actual classification agreement between the reviewers was in excess of 80% for each indicator. The final decision in cases of disagreement was reached by consensus. Shaded cells are those in which the unit decision differed from the facility decision. Examination of these tables shows that data problems restricted our ability to compute control charts at the unit level in 3 of 60 (5%) cases. Of the 57 computed unit-level control charts, only 12 (21.1%) were deemed to be non-classifiable. Of the remaining 45 charts that were classifiable, 22 (48.9%) indicated different changes in the prevalence of the condition being monitored over the reporting period than did the facility chart.

Three examples of charts with differences

Figure 1 shows the control charts for the indicator PRU05 (stages 2–4 pressure ulcer) for facility 3 and 2 of its units (1 and 4). The facility was rated as having no change in quality over 4.5 years, while unit 1 had worsening quality and unit 4 had improved quality. From mid-2007 to the end of 2011, the facility prevalence of stage 2 or greater pressure ulcers remained constant at around 3%, with some variation. Values ranged from 4% (2008 quarter 1) to 2.5% (2010 quarter 3), but performance was remarkably stable. In unit 1, however, prevalence began at 1% and slowly worsened to almost 5%. On unit 4, prevalence began at 12%, improved to 3.5% within 2 years, and then remained relatively stable at 3.5–4.5%. If those accountable for quality and safety in this facility focused solely on the facility chart, they would miss the astounding improvement on unit 4 and not respond to the potentially serious worsening prevalence rates on unit 1.

Figure 2 demonstrates performance of the indicator PAI0X (prevalence of residents with pain) in facility 13. RAI data were available for 3.25 years, 2008 quarter 4 to 2011 quarter 4. In this case, the facility measurement showed improvement from a rate of around 20% to close to 10%. Similar improvement is seen in unit 3 of this facility as shown in figure 3. In contrast, unit 4 in this facility had a rate of around 30%, with substantial variation, and showed no indication of improvement. While the facility itself appeared to be dealing better with resident pain, unit 4 had high levels of pain which did not appear to change.

Figure 4 shows the performance in facility 6 on indicator DGR01 (prevalence of residents receiving an antipsychotic with no diagnosis of psychosis). Facility 6 showed marked improvement over 4.25 years, from a rate of 40% of residents on an antipsychotic with no diagnosis of dementia to a rate of 25%. In contrast, unit 2 in this facility started with a rate of 25% but its rate finished at over 45%. If one considered only facility-level data, one might celebrate. However, examination of unit 2 data would warrant at least discussion and probably investigation to understand the rate fluctuations.

DISCUSSION

In this article, we have advanced our work on units in nursing homes and demonstrated the use of SPC as a tool for quality improvement in nursing homes. We described development of a series of explicit decision rules which can be used by experienced and naïve reviewers to classify performance in control charts. We demonstrated good inter-rater reliability for these rules

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Characteristics of the nursing home sample*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility</td>
<td>Number of beds</td>
</tr>
<tr>
<td>3</td>
<td>300+</td>
</tr>
<tr>
<td>6</td>
<td>80+</td>
</tr>
<tr>
<td>13</td>
<td>165+</td>
</tr>
<tr>
<td>17</td>
<td>150+</td>
</tr>
<tr>
<td>26</td>
<td>115+</td>
</tr>
<tr>
<td>28</td>
<td>100+</td>
</tr>
</tbody>
</table>

*The table gives approximate number of beds to assure anonymity.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Agreement in classification of control charts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility</td>
<td>Units</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>PRU05 (stages 2–4 pressure ulcers)</td>
<td>3 (50)</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>PAI0X (pain)</td>
<td>1 (33.3)</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>12 (21.1%)</td>
<td>1 (25)</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>DRG01 (antipsychotic with no diagnosis of psychosis)</td>
<td>2 (100)</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>1 (33.3)</td>
<td>1 (33.3)</td>
</tr>
<tr>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>Total 20</td>
<td>7 (35)</td>
</tr>
</tbody>
</table>

*As noted in text, we did not compute control charts for facility 17 (DRG01), thus the denominator here is 18 not 20.
and invite others to learn and use them for monitoring quality and safety in nursing homes.

Several points are of note. First, no experimental interventions were carried out in the facilities during the 4.5 years of reporting. Thus, the performance displayed in control charts reflects the natural history of these facilities and units. Second, control charts were computed on adjusted indicators so changing patient populations would have little effect on values reported. Third, in 22 of the 45 cases (48.9%), unit control charts indicated different changes in the prevalence of the condition being monitored over the reporting period than did the facility chart. Fourth, substantive differences in the interpretation of performance appear when unit-level and facility-level data are examined.

Since these indicators come from quarterly RAI–MDS 2.0 data, they are less useful for monitoring actual quality improvement processes. However, they are useful for identifying gaps in quality or safety which might be amenable to quality improvement, for demonstrating sustainability of improvement over time and for accountability purposes. We have demonstrated that, in considering only facility-level data, one may miss substantial opportunities for improvement in some units and overlook important improvements occurring on other units.

We restricted our work to three indicators on a small, but representative, sample of urban nursing homes. In this sample, we showed that unit charts display different quality behaviours than the corresponding facility charts almost 50% of the time. We did not attempt to determine how many times unit results differed from facility results for other indicators or for non-urban homes.
This may occur only infrequently, but even an infrequent difference may have an important impact on a facility's improvement and monitoring efforts.

The actual mechanics of creating control charts such as those reported in this article are quite straightforward but some technical challenges do need to be addressed. RAI–MDS 2.0 data must be identified at the unit level and units must be defined to be consistent with the definition of clinical Microsystems. Facilities must examine their definition of unit, realign these definitions as needed and ensure that the unit field is consistently completed in RAI–MDS 2.0 assessments.

This work is limited in being carried out on a small but representative sample of urban nursing homes in the Prairie Provinces of Canada. We believe these homes are representative of the Canadian context. In jurisdictions with different organisational structures, variations between unit and facility performance may not be so large. However, evidence indicates that quality improvement may be most effectively carried out in microsystems.

CONCLUSION

Our results demonstrate the necessity of considering facility-level and unit-level measurements when calculating quality indicators derived from RAI–MDS 2.0 data, and quite probably any RAI measures. We urge those accountable for collecting these data to first ensure that ‘units’ in their facility are aligned to clinical microsystems as described above, and second, ensure that unit-level identifiers are included in their datasets so that indicators can be computed at the unit level. These results support our previous work on the alignment of units in nursing homes with organisational-level microsystems. From our examples, it is clear that quality demonstrated by these important practice-sensitive indicators is a unit-level property. We believe that a focus on microsystems in nursing homes is important for system improvement.

Author affiliations
1Department of Family Medicine, University of Calgary, Calgary, Alberta, Canada
2Knowledge for Improvement Consulting, Stratford, Ontario, Canada
3Faculty of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada
4Faculty of Nursing, University of Alberta, Edmonton, Alberta, Canada
5University of Waterloo, Waterloo, Ontario, Canada
6Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
7School of Nursing, University of Ottawa, Ottawa, Ontario, Canada
8Health Quality Council (Saskatchewan), Saskatoon, Saskatchewan, Canada

Acknowledgements The authors acknowledge the contributions of the Translating Research in Elder Care (TREC) team at the time of the study: Carole A Estabrooks (Principal Investigator). Investigators: Greta G Cummings, Lesley Degner, Sue Dodson, Heather Laschingter, Kathy McGilton, Verena Menec, Debra Morgan, Peter G Norton, Joanne Profetto-McGrath, Jo Rycroft-Malone, Malcolm Smith, Norma Steward, Gary F Teare. Decision-makers: Caroline Clarke, Greta Lynn Ell, Belle Gowerluk, Sue Neville, Corrine Schalm, Donna Steinmachovich, Gina Trinidad, Juanita Treemer, Luana Whitbread. Collaborators: David Hogan, Chuck Humphrey, Michael Leiter, Charles Mather. Special advisors: Judy Birdsell, Phyllis Hempel (deceased), J Ivan (Jack) Williams, Dorothy Pringle (Chair, Scientific Advisory Committee). They assisted with recruitment of homes and local issues with respect to data acquisition.

Contributors CAE, PGN, GGC, GFT and JES contributed the recruitment of the homes, acquisition of their resident assessment instrument data and original development of our unit definition. JWP led the work in computation of unit-level and facility-level RAI quality indicators. PGN, MM, GFT and MBD conceptualised, developed and tested the decision rules for control charts and applied them to the study sample. All authors contributed to the conceptualisation and design of the article, critically reviewed the manuscript for important intellectual content and approved the final submitted version of the manuscript. CAE and PGN secured funding for the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding This work was supported by a grant-in-aid from Canadian Institutes of Health Research (MOP #53107).

Competing interests None.

Ethics approval Health Research Ethics Board, University of Alberta.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data used in this study are presently unable to be shared due to ethics and privacy restrictions. The TREC team is committed to appropriate open access of their data in the future and continues to work towards this end.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/
REFERENCES


20. CIHI. RAI-MDS 2.0 decision-support tools for clinicians and managers. Ottawa, ON: Canadian Institute for Health Information, 2012.


### Additional Table 1
Classification of the control charts for PRU05 (stage 2 to 4 pressure ulcers)

<table>
<thead>
<tr>
<th>Facility</th>
<th>Facility</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>The same</td>
<td>Worsened</td>
<td>Worsened</td>
<td>Not classifiable</td>
<td>Improved</td>
<td>Worsened</td>
<td>Not classifiable</td>
</tr>
<tr>
<td>6</td>
<td>The same</td>
<td>Not classifiable</td>
<td>The same</td>
<td>Not classifiable</td>
<td>Improved</td>
<td>Improved</td>
<td>Not classifiable</td>
</tr>
<tr>
<td>13</td>
<td>Improved</td>
<td>The same</td>
<td>Improved</td>
<td>Improved</td>
<td>Improved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>The same</td>
<td>The same</td>
<td>Worsened</td>
<td>Improved</td>
<td>Improved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Improved</td>
<td>The same</td>
<td>The same</td>
<td>Improved</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Not classifiable</td>
<td>The same</td>
<td>Not classifiable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For unit 1 of facility 1 the control chart was not computed due to insufficient numbers of appropriate RAI assessments.

### Additional Table 2
Classification of the control charts for PAI0X (pain)

<table>
<thead>
<tr>
<th>Facility</th>
<th>Facility</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Improved</td>
<td>*</td>
<td>Improved</td>
<td>Not classifiable</td>
<td>Improved</td>
<td>Improved</td>
<td>The same</td>
</tr>
<tr>
<td>6</td>
<td>Improved</td>
<td>Improved</td>
<td>The same</td>
<td>The same</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Improved</td>
<td>Not classifiable</td>
<td>Improved</td>
<td>The same</td>
<td>The same</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Worsened</td>
<td>Worsened</td>
<td>Worsened</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Not classifiable</td>
<td>Not classifiable</td>
<td>Improved</td>
<td>Worsened</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Improved</td>
<td>Not classifiable</td>
<td>Improved</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Additional Table 3
Classification of the control charts for DRG01 (antipsychotic with no diagnosis of psychosis)

<table>
<thead>
<tr>
<th>Facility</th>
<th>Facility</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Improved</td>
<td>Improved</td>
<td>Improved</td>
<td>The same</td>
<td>Improved</td>
<td>Improved</td>
<td>Not classifiable</td>
</tr>
<tr>
<td>6</td>
<td>Improved</td>
<td>Improved</td>
<td>Worsened</td>
<td>Improved</td>
<td>Improved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Improved</td>
<td>The same</td>
<td>Improved</td>
<td>The same</td>
<td>The same</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Improved</td>
<td>Improved</td>
<td>Improved</td>
<td>Improved</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>The same</td>
<td>Worsened</td>
<td>Not classifiable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary File 2

Categorizing Patterns of Change in Control Charts
Michael A. Murray, Peter G. Norton, Gary Teare
(2012)1

We used standard empirical rules commonly used in the interpretation of control charts2 to develop a set of coding rules. Our rules were then used to make a decision about whether a team’s quality indicator measurements were improving, worsening, staying the same, or were indeterminate across time. These coding rules were refined over a series of iterations. For each iteration, a new set of 20 to 25 control charts based on clinical data were prepared and scored independently by two expert reviewers. After the scoring, a consensus conference of the reviewers was held to resolve differences and the rules modified as necessary.

Desired Classification Coding

Codes for ‘predictable’ patterns
1 = better at the end
2 = worse at the end
3 = about the same at the end, and relatively stable across time
4 = about the same at the end, but with instability in the middle (codes 3 & 4 can be combined)

Code for unpredictable or unclassifiable pattern
5 = unclassifiable as to performance; erratic

NOTE: the rules were designed for the RAI results, which had relatively short series of data. For longer runs of data, consider investigating whether change happened before the end of the data series.

Specific Code Rules

Codes 1 & 2: The first two codes are concerned with the question, “Do the current (or most recent) measurements show that results3 are better or worse than where they started?”

- By ‘current’ we mean the most recent few quarters of data; we would like evidence that the process is both changed (for better or worse) and somewhat predictable at a new level for the future.
  - For longer series of data, one on the order of many years for example, we might consider whether improvement happened at any point in the series regardless of whether it was sustained. This makes for harder coding.
- We are also concerned with where the process started. Given the myriad of possible patterns in processes over time, it is possible a process could become much better or worse over time and then change again. The ‘current’ process might look better or worse than an intermediate point BUT still not be better or worse than the beginning. We are

---

2 Commonly called the “Western Electric rules” although there are many different versions of these.
3 Results could be ‘process measures’ or ‘outcome measures’; our methodology would work the same way for both.
not concerned with whether the process is now better or worse than it ever was, just whether it is better or worse than at the beginning.

- A single 3 SD rule violation either at the very beginning of the process or right at the end of the process is not sufficient evidence of a process that is better or worse at the end, or that the process is unstable and therefore unclassifiable. In combination with other nearby signals, however, it might help provide evidence of change. Especially consider discounting a widely divergent point at the beginning or end of the series in making judgements.

- In judging whether a process is better or worse at the end, it is useful to consider how many points at the beginning and end of the data series are involved in providing evidence of change.
  - At the beginning of the process we have no prior knowledge about it and would like to see evidence based on more than 1 or 2 points before deciding on the initial level of performance. Use at least 3 points (9 months when looking at RAI quarterly data) to decide. Often it appears that the first data point is wildly different than subsequent points. By itself, as noted earlier, it provides insufficient evidence to base subsequent decisions on but, if there are other rule violations at the beginning, that might be sufficient. A first data point need not be out of control to be discounted as establishing initial performance, but it should be wildly divergent from the next few values to be discounted.
  - At the end of the process, we know much more about its performance. A single divergent point at the end of the process shouldn’t be used to judge improvement or deterioration, but a 2-of-3 rule violation might be sufficient. The two other multiple point rules (4-of-5 or 8-on-a-side) are better evidence of sustained change.

- When viewing a process with relatively stable beginning and ending periods but instability in the middle, be wary of considering a process as having changed (up or down) when mean differences are small. One of the longer multi-point rule violations (4-of-5 or 8-on-a-side) at the beginning or end might be sufficient to say the process is better or worse, but just seeing slightly lower or higher scores is insufficient evidence of change.

- Trend rules are hard to use when evaluating processes. A trend violation at the beginning of the data series says that the process isn’t stable to begin with and deciding the actual starting level of the process then becomes hard. At the end, a trend might indicate improvement or decline BUT what matters is where the trend ends. A process could be improving but still be worse than where it started, or vice versa. Trend violations are unlikely, therefore, to be good measures of improvement or decline, but may, in combination with other rules, give some evidence of change.

Codes 3 & 4: The third and fourth codes are concerned with the question, “Given that the current scores look to be at about the same level as where they started (i.e. not coded 1 or 2), were the results relatively stable across time or did they show substantial variation in the middle?”

- A ‘pure’ code 3 pattern would show a process with no rule violations over its entire length.
• A relatively stable process might have a rule violation that is inconsequential in understanding long term performance. It might be a 3 SD violation in the middle, or a trend, or an isolated 2-of-3 violation. Consider this a code 3.
• The current process could have roughly the same level as the beginning but with substantial variation in between. Multiple rule violations, especially when they are on both sides of the centre line, are probably evidence of a code 4.

Code 5: A process might be sufficiently erratic that deciding whether it has improved, worsened or stayed the same is just not possible. A process might be just too erratic over the whole series to decide what is happening, or with too few stable points at the end to be confident about its most recent performance. In that case, consider using code 5.

• The last few subgroup values might not be stable enough to judge how the process is actually performing and to get any idea about whether it is better or not.
• There might be multiple rule violations on both sides of the centre line from beginning to end, and the process may appear to be just not stable.
• Deciding on whether a process is ‘about the same but with instability’ or ‘unclassifiable’ may be difficult. The distinction is based on how many points are available to judge the ending of the series. If a period of higher or lower performance in the middle or near the end of the process gives insufficient points to be confident in a judgement or statement that a process has changed, then consider the process ‘unclassifiable’.

General coding considerations and issues
• If you find yourself waffling between codes, especially a code 4 or something else, consider using code 5.
• For research purposes, you might want to be conservative in saying a process is better or worse, erring on the side of codes 3 through 5.
• When considering giving feedback to organizations, these codes and rules might be too conservative and not give sufficient credit for improvements over the medium term. Yes, performance might not have changed over the long term, but over the most recent year or so it might have.
• Given the relatively short data series available and the way control limits are calculated using Moving Ranges, a single HUGE change in values might broaden limits and change the averages. The Nelson method compensates for this and you might want to consider running a few charts through it. As a shortcut, when you see a subgroup value that is wildly different from all others, consider whether some values might become a violation as the limits shrink slightly AND change your classification. We can investigate those cases in subsequent analyses.

---

4 In an XmR (individuals) chart, the Nelson rule removes points with a 3 SD violation ON THE MOVING RANGE (mR) CHART from calculation of limits on the X chart. Huge jumps in points, which affects average moving range, might cause points to be removed from calculation of limits; it isn’t just about high or low points.
Examples to show coding rules

Charts show 3 SD limits & 3 SD violations as well as other rules.

1=Better

Starts with 2 out-of-control high points (1 wouldn’t have been enough and even 2 is suspect), but ends with 6 zeros.
1=Better

Just barely worse because the final 2 points are out-of-control. One point alone at the end would have been insufficient evidence of worsening. Using a Nelson rule would probably NOT have been enough to cause the first 2 points to violate a rule.

2= Worse

Just barely worse because the final 2 points are out-of-control. One point alone at the end would have been insufficient evidence of worsening. Using a Nelson rule would probably NOT have been enough to cause the first 2 points to violate a rule.
2=Worse
Starts low, ends high (4 of last 5 above 1 sigma)
2=Worse

Ignore first point and spike in last point. The worsening trend in the middle is NOT sustained long enough.

3=Same

Ignore first point and spike in last point. The worsening trend in the middle is NOT sustained long enough.
3=Same
No rule violations. Re-running with the Nelson rule would probably push the first point and fourth-last point out of control, but the process would probably still be stable.
3=Same

The last 4 points are not sufficiently higher than the first few to say with confidence that the process is worse. A further 1 or 2 points at the same level might say so.

4=Same with instability

The last 4 points are not sufficiently higher than the first few to say with confidence that the process is worse. A further 1 or 2 points at the same level might say so.
4=Same with instability

The first two points are higher, then the next 5 are about the same level, but 2 more points show that the process is better. The process then goes back to the middle, then high (which might be out-of-control using the Nelson rule), then back down. It is hard to predict the next point and impossible to say if the process is better or worse since its start. The last high is worse compared to most of the middle half.

5=Unclassifiable

The first two points are higher, then the next 5 are about the same level, but 2 more points show that the process is better. The process then goes back to the middle, then high (which might be out-of-control using the Nelson rule), then back down. It is hard to predict the next point and impossible to say if the process is better or worse since its start. The last high is worse compared to most of the middle half.
5=Unclassifiable
Although the process starts off at a good level, it is worse in the middle and the final 6 points are confusing. Whether the process will continue higher (and therefore worse) or back down to the third- and fourth-last points is unpredictable.

![Control Chart: Raw PRU 2+](chart1)

5=Unclassifiable
Technically this is a code of 3 because there are no rule violations, but application of the Nelson rule would push the high point out. Absolute differences in performance are very large. The start and end points are zero but clearly the final few points aren’t zero just as the first 4 points aren’t zero.

![Control Chart: Raw PRU 2+](chart2)