LASA-ACSA submission on star ratings

Tim Hicks, General Manager, Policy, Advocacy and Advisory, Leading Age Services Australia

December 2021

Background

Government is consulting on star ratings in residential care to be introduced by 1 January 2023, and a broader quality indicator program for which implementation details are still being determined.

The proposed star rating system is to be informed by care minutes, Consumer Experience Interview scores with sample of clients, scores on quality indicators and accreditation performance based on the existing rating scale.

Meaningful differences not forced ranking

Our view is that the purpose of the star ratings is to provide consumers with insight into the quality of a service. This means star ratings should not be a forced ranking exercise where providers are scored based on where they fit within a distribution regardless of whether this difference reflects practically meaningful differences in quality. In other words, star ratings should be based on agreed criteria across the broad groupings and providers then allocated a star rating depending how they score in each of the weighted groupings if an overall rating is produced.

Continuous improvement

Given the timeframes, what can be implemented in time for 1 January 2023 will be only a minimum viable product. It is therefore essential to build a plan for continuous improvement into the implementation timetable. This continuous improvement work should be overseen by an independent expert group.

Overall score

Many providers believe that there should not be an overall weighting at all given the arbitrariness of summing the different indicators.

If the Government chooses to proceed with an overall rating this should be the sum of the sub-indexes multiplied by their weightings.

Consumer experience

Most providers and consumers that we have engaged with over the last several years strongly support a heavy weighting being given to consumer experience scores. However, there are some methodological issues that need to be resolved with the proposed methodology for consumer experience.

- Assuming methodological issues can be resolved, we recommend that the consumer experience sub-index count for 50% the overall score.
- We recommend that the following methodological issues be resolved:
    - The sample size needs to be determined based on power analysis to identify differences of a magnitude of X (TBD) with Y (TBD) tolerance for type 1 error (failing to identify a difference that does exist) and Z (TBD) tolerance for type 2 error (falsely identifying a difference that does not actually exist).
        - Sampling methodology also needs to be appropriately randomised. Extra care must be taken to avoid bias in only sampling residents more likely to be responsive to requests.

        - Where there is a mix of cognitively impaired and non-cognitively impaired residents the sample should match the population distribution.
    - Where outlier results are found, consideration should be given to increasing sample size to confirm the accuracy of the finding.
    - More regular than yearly reporting should also be considered to improve reliability and provide more frequent opportunities for services to improve.
    - Results need to be risk adjusted based on respondent characteristics. Most of the data required to do this is likely to be available through existing sources such as AN-ACC. However, there may be some additional data required as indicated below.
        - We know that family members generally give lower ratings than residents. Family members tend to have less knowledge of the service than residents. Some providers consider that it would be more accurate to only sample residents because the views of non-cognitively impaired residents are closer to the experience of other cognitively impaired residents than family members with limited knowledge of the service. However, other providers think that is important for cognitively impaired residents to have someone speak on their behalf, and seek to address the issue of lower ratings through risk adjustment. It is broadly agreed that family members should only be included in the sample as representatives or residents that are too cognitively impaired to complete the instrument. Versions of the instrument for people with mild to moderate cognitive impairments should be used where possible. A response to situations where there is significant divergence between resident and representative responses needs to be considered.
        - Resident need and healthcare status needs to be included in the model so that risk adjustments can occur if any of these variables are correlated with consumer experience scores. Much of this data should be available through AN-ACC.
        - Resident / representative demographics should also be considered. Anecdotally, affluent consumers have higher expectations and will give lower scores. Some measure of affluence needs to be considered. This could be means testing data or if this is not available something like the ABS Index of Relative Socio-Economic Advantage and Disadvantage for the region. However, there is some concern that this may be contrary to principles of equity.
        - As part of the trial, additional questions should be included in the instrument to act as potential risk adjustments. For example, the instrument may ask about the person's consumer experience in other common service contexts, or even a self-assessment about whether the person considers themselves to have higher than average standards.
        - Seasonal adjustment or more regular collection may also be required.
    - Scoring individual questions based on how many people respond with the lower two answers in the rating scale (as seems to have been done historically) throws away significant information, and the basis for this needs to be explained.
    - The underlying constructs for the instrument need to be explained as many of the questions appear to go to very similar issues whereas other are more distinct. This could include undertaking item discrimination analysis.
    - Relatedly, rules for validating responses needs to be determined and explained. E.g., a response saying that staff are never kind and caring but always respectful should be regarded sceptically.

    - The answer scales need to be revisited. Currently, taken literally, the responses map to frequencies of 0% (never) 1-49% (sometimes) 50-99% (mostly), 100% (always). It seems the response to this has previously been to group never and sometimes together with mostly and always. But apart from throwing away information, this still means people between the categories may be closer than people within the categories. Frequency based scales also fail to distinguish a relatively minor one-off failure from a major one-off failure. There is a case for avoiding emotional intensity-based scales (e.g., extremely satisfied to extremely dissatisfied) because they elicit more subjective responses. But a better solution than the current scale is needed.
    - Depending on how the rating scale issue is addressed, consideration needs to be given to what constitutes a good or bad score within each question, as well as for the overall instrument. For example, a person occasionally having a meal they don't like is not surprising. Whereas a person occasionally not receiving the care they need would be more serious.
    - These issues should ideally be addressed through a formal validation study, including comparing the scores for the revised CER instrument to other instruments.
- As part of the resolution of the above methodological issues, a scoring system needs to be determined. Conceptually, we recommend that that the scoring system be developed to reflect whether the resident is judged to have had a good experience.
- The cut points for stars should then be based on the likelihood of a person having a good experience, based on feedback about what constitutes meaningful differences along this scale.
- For illustrative purposes – noting the method will need to depend on resolution of issues such as the rating scale and definition of underlying constructs – this could look as follows:
    - Each of the items in the current scale could be assigned a rank from 1 to 4. This means the maximum score across, assuming there continue to be 10 questions, would be 40 while the minimum score would be 10. A good experience could be defined as 30, reflecting an average rating of at least most of the time.
    - Cut points of for the ratings could be defined as
        - One star = <50% likely to have a good experience
        - Two stars = 50-69% likely to have a good experience
        - Three stars = 70-79% Likely to have a good experience
        - Four stars = 80-89% likely to have a good experience
        - Five stars = 90-99% likely to have a good experience
- Knowledge of the distribution of scores could inform the initial discussion about what constitutes a meaningful difference in likelihood of having a good experience. However, basing cut points on the distribution is problematic because (a) the distribution changes and this discourages investment in improving quality because it makes it harder to predict whether efforts to improve will result in a tangible improvement in score (b) differences in ranking within a distribution may not be practically meaningful, and in the case of consumer experience the fact that results cluster around very positive scores mean that a provider could have an objectively good score but still be relatively low within the distribution.

Staffing

Staffing is an important input to care. Including input indicators in the rating scale helps to guard against potential issues with the comprehensiveness or risk weighting of outcome indicators.

- Staffing should be weighted at 20%

- A strong view for many is that it is important that staff quality, and allied health and lifestyle staff are taken into account in the staff rating
- One approach to setting the cut points would be as follows:
    - One star should reflect a service that fell significantly below the mandated minimum staffing and did not qualify for an exemption.
    - Two stars should reflect a service that is no more than 5 minutes below the mandated minimum staffing levels.
    - Three stars should reflect a service that delivers the mandated minimum staffing levels but does not have enough staff to meet the four-star criteria
    - For four-star and five-star levels staffing scores should reflect average hours of nursing, personal care, lifestyle staff, and allied health, weighted by staff pay rates to reflect different levels of skill and experience (some consideration on how to deal with agency staffing is required). Including wider staffing in the determination of four-star and five-star levels helps to address some of the perverse incentives from not including these staff in the minimum staffing rules. Further consideration of the specific benchmarks is needed.
- An alternative view within industry is that if compliance is rated at five stars for accreditation, it should also be rated at five stars for staffing. The concern is that it will be confusing for consumers to understand the different between the scales. Whereas on most of the scales five stars is some indication of excellence, for the compliance scale it reflects meeting minimum standards.

Compliance

Notwithstanding current sector concerns about the standards and the way they are enforced, compliance scores provide an important process indicator.

- Compliance scores should be weighted at 20% noting that the vast majority of services will be fully compliant most of the time.
- The sector supports the current rating scale for compliance. However, systems need to be improved so that returns to compliance are more quickly updated.
- Many in the sector support a move to UK style graded assessments in the future, but there is no confidence that current standards and audit methodologies are robust enough to support this.

Clinical indicators

Clinical indicators capture important dimensions of care. A small number of adverse clinical events can be a serious issue even in the context of high consumer satisfaction. However, we have a number of significant concerns with the chosen indicators.

- We recommend that clinical indicator scores should be weighted at only 10% of the total score, with major problems with care quality instead identified through regulatory assessments until methodological problems can be resolved.
- To smooth chance variation in scores, the results incorporated into the index should be a rolling 12-month average.
- Clinical outcomes are a function of a person's condition and choices as well as the care offered by the provider – this needs to be effectively adjusted for and we are concerned that this cannot be done well at a service-wide level.

- Some indicators, such as falls, also reflect a trade-off between safety and dignity of risk. It is not clear that this can be adjusted for
- The self-collection of clinical indicator data is also a fundamental limitation in the usefulness of this data for public reporting. Consideration needs to be given to methodologies for checking and/or auditing reported data to ensure a level playing field. For example, checking that the number of residents in which data is submitted on are actually the number of residents in a facility. Consideration should also be given to the use of administrative data as per the ROSA Outcome Monitoring System.
- A process needs to be identified to discuss and determine what should constitute a good score. This needs to be informed by access to risk adjusted data on the distribution of scores between facilities and over time.
- This is complex and necessarily subjective process and needs to take into account not just the frequency of the outcomes but also the relative seriousness – for example stage 4 pressure injury prevalence should be weighted more heavily than stage 2 prevalence.
- One option might be to take the events being tracked and assign a seriousness score to each event. The post risk adjustment probability of an event occurring, weighted by the events seriousness could then be used to provide a total score. This allows the different indicator data to be pooled into a single measure of adverse clinical events.
- As with consumer experience indicators, cut points should then be assigned based on meaningful differences in the probability of having an event of some sort occur. Looking at data published by the Royal Commission the interquartile range (i.e., 25th to 75th percentile) tends to be 5 to 10 instances per 100 residents for high prevalence events such as falls, pressure injuries, and unplanned weight loss. With median rates in range of roughly 5 to 15 events per hundred residents for each indicator.
    - Again, we are sceptical that data can be risk adjusted well enough to avoid creating incentives to not accept residents likely to bring down a facility's score.