BMJ Open

# Testing quality indicators and proposing benchmarks for physician-staffed emergency medical services: a prospective Nordic multicentre study

Helge Haugland ,[1,2] Anna Olkinuora,[3] Leif Rognås ,[4,5] David Ohlen,[6] Andreas Krüger[1,2]

**To cite:** Haugland H, Olkinuora A, Rognås L, *et al*. Testing quality indicators and proposing benchmarks for physician-staffed emergency medical services: a prospective Nordic multicentre study. *BMJ Open* 2019;**9**:e030626. doi:10.1136/bmjopen-2019-030626

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-030626).

Received 25 March 2019
Revised 18 September 2019
Accepted 11 October 2019



For numbered affiliations see end of article.

**Correspondence to**
Dr Helge Haugland;
helge.haugland@norskluftambulanse.no

## ABSTRACT

**Objectives** A consensus study from 2017 developed 15 response-specific quality indicators (QIs) for physician-staffed emergency medical services (P-EMS). The aim of this study was to test these QIs for important characteristics in a real clinical setting. These characteristics were feasibility, rankability, variability, actionability and documentation. We further aimed to propose benchmarks for future quality measurements in P-EMS.

**Design** In this prospective observational study, physician-staffed helicopter emergency services registered data for the 15 QIs. The feasibility of the QIs was assessed based on the comments of the recording physicians. The other four QI characteristics were assessed by the authors. Benchmarks were proposed based on the quartiles in the dataset.

**Setting** Nordic physician-staffed helicopter emergency medical services.

**Participants** 16 physician-staffed helicopter emergency services in Finland, Sweden, Denmark and Norway.

**Results** The dataset consists of 5638 requests to the participating P-EMSs. There were 2814 requests resulting in completed responses with patient contact. All QIs were feasible to obtain. The variability of 14 out of 15 QIs was adequate. Rankability was adequate for all QIs. Actionability was assessed as being adequate for 10 QIs. Documentation was adequate for 14 QIs. Benchmarks for all QIs were proposed.

**Conclusions** All 15 QIs seem possible to use in everyday quality measurement and improvement. However, it seems reasonable to not analyse the QI 'Adverse Events' with a strictly quantitative approach because of a low rate of adverse events. Rather, this QI should be used to identify adverse events so that they can be analysed as sentinel events. The actionability of the QIs 'Able to respond immediately when alarmed', 'Time to arrival of P-EMS', 'Time to preferred destination', 'Provision of advanced treatment' and 'Significant logistical contribution' was assessed as being poor. Benchmarks for the QIs and a total quality score are proposed for future quality measurements.

## Strengths and limitations of this study

- This is the first study putting the EQUIPE (Establishing Quality Indicators in Physician-staffed Emergency Medical Services) quality indicators (QIs), developed specifically for physician-staffed emergency medical services, into a clinical setting.
- A prospective multicentre study involving 16 Nordic physician-staffed helicopter emergency medical services.
- The QIs are assessed for important QI characteristics.
- Benchmarks for future quality measurement are proposed.
- Except from the feasibility of the QIs, the assessment of the different QI characteristics was done by the author group.

## INTRODUCTION

### Background/rationale

The importance of quality improvement in healthcare has been recognised by leading health organisations and in landmark publications.[1–4] However, publications on quality measurement in physician-staffed emergency medical services (P-EMS) are rare.[5] For prehospital services in general, and P-EMS specifically, more research on quality measurement has been warranted.[6 7] Moreover, it has been argued that quality assurance and even quality improvement in P-EMS requires a model for quality estimation to achieve appropriate governance.[8] Quality measurements are an obvious prerequisite for quality improvement. A first initial step is the development of appropriate tools for quality measurement, that is, quality indicators (QIs). A QI can be defined as a measurable element of performance for which there is evidence or consensus that it can be used to assess the quality and hence change the quality of care provided.[9]

No comprehensive set of systematically developed QIs are registered in P-EMS in Sweden, Denmark, Finland and Norway. Attempts on extracting information concerning the quality of the service have

primarily been limited to time variables.[10] Response time has been widely used for quality assessment but may have been overemphasised and is not applicable for all prehospital emergency medical activity.[11] Time variables primarily describe the transport component of P-EMS. This information is necessary but not sufficient for quality assessment. The care component of P-EMS also has to be addressed. In fact, The Institute of Medicine, a US independent non-governmental research organisation, has defined six quality dimensions that should be addressed when measuring the overall quality of a health service[12]: patient centredness, safety, effectiveness, efficiency, equity and timeliness. If only one or a few of these quality dimensions are addressed, the result can be a simplistic and narrow quality measurement.

In 2018, we published a systematic literature review describing quality measurement studies in P-EMS.[5] There was no common understanding in the studies as to which QIs to use. Moreover, 15 out of the 27 identified studies used only one QI. This increases the risk of a one-sided approach in quality measurement. The review concludes that future quality measurement in P-EMS should be done based on a consensus-based set of QIs rather than a single QI to ensure a comprehensive quality measurement. In another recent study, we developed a set of multidimensional QIs for P-EMS through a consensus process. These QIs were called the EQUIPE (Establishing Quality Indicators in Physician-staffed Emergency Medical Services) QIs (online supplementary file 1). Panellists from different stakeholder groups agreed on 15 response-specific QIs for P-EMS.[13] These are QIs that should be feasible to collect from any P-EMS response during the prehospital time interval or in the emergency department at handover. Despite methodically correct development, QIs are not necessarily suitable in real datasets. The actual QIs have not yet been tested in clinical datasets. Based on modern framework for QI efforts, the next stage in the development of QIs for P-EMS should be testing for critical QI characteristics (feasibility, rankability, variability, actionability and documentation).

### Objectives

The aim of this study was to test the multidimensional QIs for the above-mentioned characteristics in a real clinical setting. We further aimed to propose benchmarks for future quality measurement in P-EMS based on the data in this study.

## METHODS

### Study design and setting

In this prospective observational study, 16 physician-staffed helicopter emergency services in Finland, Sweden, Denmark and Norway registered data for the EQUIPE quality indicators. There has previously been documented significant system similarities in the P-EMS of the four participating countries, making them a suitable arena for multicentre studies.[14] The Nordic countries have a mix of urban of rural areas with a rather low overall population density (19.6 inhabitants/km$^2$). The prehospital incidence of critical illness and injury in these countries has been documented to be 25–30/10 000 person-years.[15] The physicians staffing Nordic P-EMS are usually experienced anaesthesiologists, most of them working both in P-EMS and in hospitals.[14 16] All Nordic services do primary responses, and the Swedish, Danish and Norwegian services also do secondary responses; the former is defined as responses where the patient is located outside a hospital, and the latter is interhospital transfers. Moreover, the Norwegian services also do search and rescue responses (SAR responses). In addition, one Swedish (Karlstad) and all Finnish and Norwegian bases dispose a rapid response car for responses close to the base and for responses in poor weather conditions that prevent flight operations. The study applied Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[17]

### Inclusion criteria and data variables

We included every request to the P-EMS to dispatch the P-EMS unit. Thus, we could include both completed and cancelled responses, as well as stand-downs (responses cancelled by dispatch or crews on-scene) and rejected responses. Examples of reasons for rejecting a response might be weather conditions or the lack of medical need as judged by the P-EMS physician. The latter is possible in Sweden, Finland and Norway where the acceptance or rejection of a response is at the P-EMS physicians' discretion. Inquiries with counselling as the only purpose were excluded. Primary and secondary responses as well as SAR responses were included. For bases with both a helicopter and a rapid response car, responses were included regardless of the mode of transportation. All 15 EQUIPE QIs were registered in responses involving patient contact.[13] Only 4 of the 15 QIs were registered in responses not involving patient contact (QIs 1, 6, 7, 10). Data were collected for 3 months (from 10 June to 12 September 2016).

### Data sources/measurement

Finland collected the necessary data by including the QIs as part of their existing documentation database (FinnHEMS database, FHDB). FHDB is a national database, including both response and patient data where all HEMS units register all responses. Some QIs could be gathered from the existing data (eg, time stamps) and those that could not were implemented either as permanent variables or on a separate study sheet. It was mandatory to fill in all the QIs in the system. The other nations registered the same data by using a web-based questionnaire (Formsite; Vroman Systems, Chicago, Illinois, USA). In all nations, the data were collected after completed response by the P-EMS physician. The four national investigators monitored the documentation of participating P-EMS bases to secure accurate data collection.

The first 2 weeks of the data collection period (from 10 June to 24 June 2016) was a feasibility test; we wanted to study if the QIs from the consensus process were feasible to collect in the everyday of P-EMS. The feasibility test was done as a pilot study involving the same Finnish, Swedish and Danish bases that participated in the main study. However, only two Norwegian bases participated in this pilot study (Trondheim and Ørland). We considered this sample sufficient because feasibility tests can be run in a small scale.[18] Here, all the recording physicians could comment on the feasibility of obtaining the necessary data. An assessment of the feasibility of the QIs was done after these 2 weeks. This was done based on comments from the recording physicians. After these 2 weeks of feasibility testing, we adapted and clarified the wording of some QIs and then continued the data collection for a total of 3 months.

We assessed four other important characteristics of QIs in addition to feasibility: rankability, variability, actionability and documentation.[19 20] This was done according to the criteria for good QIs defined by the Organisation for Economic Cooperation and Development and the Agency for Healthcare Research and Quality.

Rankability is assessed by judging if a QI has a clear direction of good and bad, that is, the QI has a good rankability if high values for a QI are always better than low values. Conversely, rankability is poor if high values are better than low values but *very* high values are worse than low values.

According to criteria for QIs, a good QI must have enough variability to allow for improvement. To assess variability, we calculated the mean and median as well as the corresponding variance for each of the QIs based on the data collected after the feasibility test. This illustrates both the average performance and the variation in the participating Nordic P-EMSs. To the best of our knowledge, there is no definition of how much variability a QI should have to be useful. This implies that the assessment of variance is somewhat arbitrary.

Actionability is the possibility of influencing the QI performance. For instance, a P-EMS has limited opportunity to reduce the time to definitive care because this mainly depends on the distances that the P-EMS unit has to work with. In that case, actionability is rather low.

Furthermore, for a QI to be valid, the process or structure of defining the QI must have been documented to give better outcome. The degree of such documentation was assessed for each QI.

We do not report which results belong to the specific P-EMS bases simply because the aim of this study was to assess the characteristics of the QIs and not to compare the performance of the participating services.

### Missing data

Due to technical solutions, the QIs 'P-EMS involvement in dispatch' and 'Debriefed responses' were registered only in responses with patient contact in Finland; however, these QIs were registered for all responses in the other three nations. The proportion of missing data for the QIs varied between 0.2% and 0.9%. Missing observations were acknowledged and omitted from the analysis. All analyses were done on variables present, thus minimising information loss.

### Statistical methods

Descriptive statistics are reported. The QI proportions were recorded for QIs that are categorical variables; time was recorded in minutes for QIs that were continuous time variables. All QIs are reported by the mean and the corresponding 95% CI as well as the median with corresponding IQR.

We also used figures from the 16 P-EMS bases to propose benchmarks for all QIs. We set the benchmark at the lower end of the fourth quartile for QIs where higher values reflect better performance. For QIs where lower values reflect better performance, we have set the benchmark at the highest end of the first quartile. We depicted the benchmarking graphically so that performances within the IQR are shown in yellow. Performances better than the IQR level are in green, and those worse than the IQR level are red.

### Ethics approval and consent to participate

According to the approvals from all four countries, the data were obtained without informed consent from patients or their next-of-kin. As stated in the study protocol, there was no deviation from regular clinical practice during the study period.

### Patient and public involvement

The QIs used in this study were developed by an expert panel through a consensus process.[13] One of the 18 members of the expert panel was a leader from a leading Norwegian patient organisation. This was done to secure user-expertise in the development of QIs.

For this particular study, no patients were involved in setting the research question, nor were they involved in the design or conduct of the study. No patients were asked to advise on the interpretation or writing up of results. The results will be disseminated via our local authorities and conference presentations. There are no plans to disseminate the results of the research to study participants.

## RESULTS

Despite the thorough and explicit definitions of QIs, a feasibility test was done first because this generally identifies variables that require modification. Omitting the feasibility test is not recommended.[18] Based on the experiences and comments from both recording physicians and the national coordinators during the 2 weeks feasibility test, we concluded that the necessary input data for the QIs were available in the participating services. There was no feedback indicating that the data were difficult to obtain. However, the definition of four QIs required

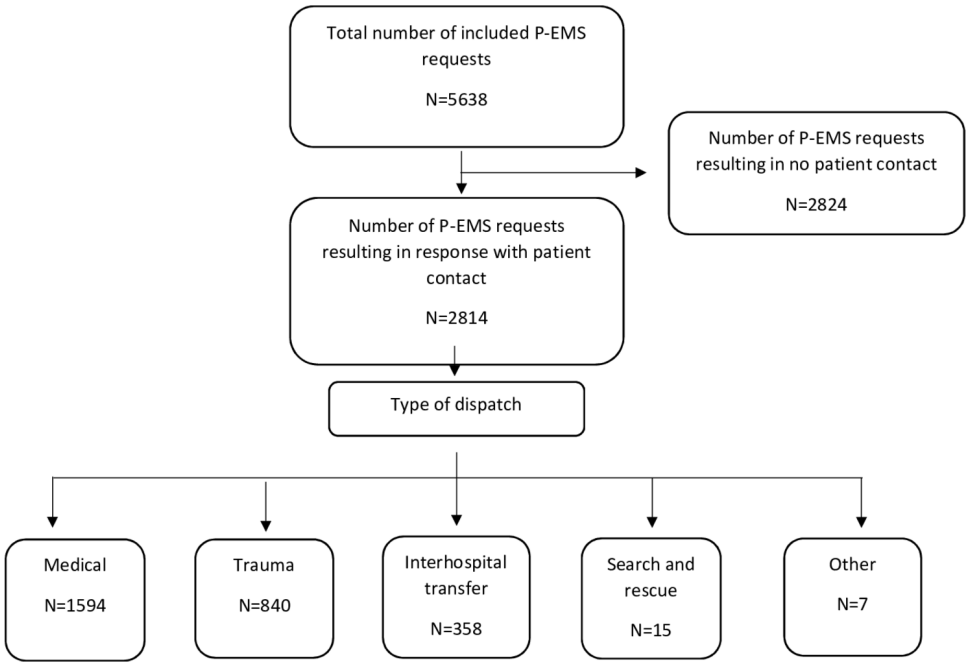


**Figure 1** Study population with type of dispatch for physician-staffed emergency medical services (P-EMS) responses with patient contact.

clarification. The changes done by the study group are documented in online supplementary file 2.

### Participants and descriptive data

The dataset consists of 5638 requests for P-EMS. There were 2814 requests that resulted in completed responses with patient contact. Reasons for requests without patient contact may be cancelled responses, rejected responses due to weather or no need for P-EMS as judged by the P-EMS physician. The different dispatch types for the responses with patient contact are depicted in figure 1.

### Outcome data and main results

The assessment of the QI feasibility, variability, rankability, actionability and documentation is depicted in table 1. The feasibility assessment was done based on comments from the recording physicians. The other four QI characteristics were assessed by the authors. The variability assessment of the QIs was based on the figures in table 2; the base-specific mean and median values with corresponding variances are shown for each QI. Documentation was assessed based on the existing literature.

**Table 1** Essential characteristics of the applied quality indicators

| Quality indicator | Feasibility | Rankability | Variability | Actionability | Documentation |
|---|---|---|---|---|---|
| Able to respond immediately when alarmed | Good | Good | Good | Poor | Fair[28 29] |
| Time to arrival of P-EMS | Good | Good | Good | Poor | Fair[28 29] |
| On scene time | Good | Fair | Fair | Good | Fair[11 30–32] |
| Time to preferred destination | Good | Good | Good | Poor | Good[33 34] |
| Patients arriving hospital alive | Good | Good | Fair | Fair | Good[35 36] |
| Debriefed responses | Good | Good | Good | Good | Fair[37 38] |
| Adverse events | Good | Good | Poor | Good | Good[39 40] |
| Complete documentation | Good | Good | Good | Good | Good[41 42] |
| Guidelines for actual medical problem | Good | Good | Good | Good | Fair[43–46] |
| P-EMS involvement in dispatch | Good | Good | Good | Fair | Poor[47] |
| P-EMS necessary to provide appropriate care | Good | Good | Good | Fair | Fair[48 49] |
| Provision of advanced treatment | Good | Good | Good | Poor | Fair[50 51] |
| Significant logistical contribution | Good | Good | Good | Poor | Good[33 34 52] |
| Patients enrolled in research projects | Good | Good | Fair | Good | Fair[7] |
| Care for relatives | Good | Good | Fair | Good | Fair[53–55] |

P-EMS, physician-staffed emergency medical services.

Table 2 Variability of QIs (note: the columns 'minimum mean value' and 'maximum mean value' refer to the lowest and highest mean values from the participating P-EMS bases)

| QI | No. of responses included | Missing (N) | Unit of QI | Mean (95% CI) | Median (IQR) | Minimum mean value | Maximum mean value |
|---|---|---|---|---|---|---|---|
| Able to respond immediately when alarmed | 5599 | 39 | % | 89 (86 to 92) | 90 (84–94) | 78 | 97 |
| Time to arrival of P-EMS | 2428 | 6 | minutes | 27 (24 to 30) | 26 (23–31) | 18 | 36 |
| On scene time | 2427 | 7 | minutes | 20 (19 to 22) | 21 (19–22) | 14 | 26 |
| Time to preferred destination | 2226 | 19 | minutes | 63 (59 to 67) | 63 (58–69) | 46 | 74 |
| Patients arriving hospital alive | 2809 | 5 | % | 91 (89 to 93) | 92 (88–94) | 85 | 98 |
| Debriefed responses | 2809 | 5 | % | 74 (64 to 83) | 78 (64–88) | 29 | 97 |
| Adverse events | 5572 | 27 | % | 2 (1 to 3) | 1 (1–3) | 1 | 7 |
| Complete documentation | 2798 | 16 | % | 64 (51 to 76) | 76 (34–80) | 25 | 91 |
| Guidelines for actual medical problem | 2802 | 12 | % | 60 (48 to 72) | 64 (45–77) | 15 | 87 |
| P-EMS involvement in dispatch | 3669 | 29 | % | 47 (27 to 66) | 34 (12–94) | 7 | 98 |
| P-EMS necessary to provide appropriate care | 2808 | 6 | % | 39 (35 to 43) | 39 (34–43) | 27 | 52 |
| Provision of advanced treatment | 2804 | 10 | % | 49 (43 to 55) | 48 (39–58) | 33 | 71 |
| Significant logistical contribution | 2795 | 19 | % | 43 (32 to 55) | 51 (24–58) | 6 | 80 |
| Patients enrolled in research projects | 2788 | 26 | % | 6 (–1 to 13) | 0 (1–3) | 0 | 40 |
| Care for relatives | 2803 | 11 | % | 94 (92 to 96) | 94 (93–97) | 87 | 100 |

P-EMS, physician-staffed emergency medical services; QI, quality indicator.

Actionability was assessed as adequate for 10 QIs. The actionability of the QI 'Able to respond immediately when alarmed' was assessed as being poor because this is primarily determined by weather and concurrency conflicts. Further, the actionability was assessed as being poor for the QIs 'Time to arrival of P-EMS' and 'Time to preferred destination' because these time variables largely depend on where the patient is located geographically, and the P-EMS service cannot influence this. Moreover, the actionability was assessed as being poor for the QIs 'Provision of advanced treatment' and 'Significant logistical contribution'. In our opinion, this is primarily the case for P-EMS services who are not involved in the dispatch decision. The actionability of these two QIs is fair in P-EMS services where the acceptance of a request is at the P-EMS physician's discretion.

We used the data from the participating bases as a description of the current performance status pertaining to the QIs. Based on these figures, we proposed a benchmark level and a graphical presentation of three performance levels for the different QIs. Yellow area represents average performance, red represents low performance and green is high performance. Our objective was that these benchmarks serve as a tool for quality improvement in comparable P-EMSs in the future. The benchmarking is presented in figure 2.

Table 3 shows how the benchmarking system can compare the performance of different bases. In the actual example, we used two of the participating bases as examples and call them Base 1 and Base 2. In the table, the actual value for each QI and its corresponding benchmark colour is depicted for all 15 QIs. For every high performance, the bases are given one point. For every low performance, the bases are given −1 point. The average performances are given 0 point. Thus, we end up with a sum or a total quality score that is between −15 and 15 for each base.

## DISCUSSION

### Key results

A set of 15 QIs were developed by an expert panel for P-EMS and were tested by applying the QIs in 5638 responses from 16 Nordic P-EMS bases. The feasibility of obtaining the necessary data for these QIs was good. The variability of the QIs was evaluated and is acceptable for all QIs except from the QI 'Adverse events'. We used the dataset to propose benchmarks for all QIs as well as a total quality score: both of these can be used as tools for future quality measurement in P-EMS. Nonetheless, we assessed the actionability of some QIs to be low. That is especially true for QIs that measure the timeliness of P-EMS.

### Interpretation and generalisability

The patients treated by Nordic P-EMS services are heterogeneous: primary trauma and medical responses for every age group, secondary transports including neonatal

| Quality indicator | Unit of QI | Performance assessment |
|---|---|---|
| Able to respond immediately when alarmed | % | 75 80 85 90 95 100 |
| Time to arrival of P-EMS | minutes | 0 10 20 30 40 50 |
| On scene time | minutes | 0 10 20 30 40 |
| Time to preferred destination | minutes | 0 20 40 60 80 |
| Patients arriving hospital alive | % | 80 85 90 95 100 |
| Debriefed responses | % | 0 20 40 60 80 100 |
| Adverse events | % | 0 2 4 6 8 10 |
| Complete documentation | % | 0 20 40 60 80 100 |
| Guidelines for actual medical problem | % | 0 20 40 60 80 100 |
| P-EMS involvement in dispatch | % | 0 20 40 60 80 100 |
| P-EMS necessary to provide appropriate care | % | 0 20 40 60 80 100 |
| Provision of advanced treatment | % | 0 20 40 60 80 100 |
| Significant logistical contribution | % | 0 20 40 60 80 100 |
| Patients enrolled in research projects | % | 0 2 4 6 8 10 |
| Care for relatives | % | 86 88 90 92 94 96 98 100 |

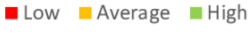


**Figure 2** Benchmarking of quality indicators (QIs). Green zone, high performance; yellow zone, average performancel; red zone, low performance. The benchmark is set at the transition between green and yellow zones and marked with a black and fat vertical line.

Table 3 Illustration of comparison between services using the proposed benchmarks

| | | P-EMS base | |
|---|---|---|---|
| Quality indicator | Unit of QI | Base 1 | Base 2 |
| Able to respond immediately when alarmed | % | 96 | 83 |
| Time to arrival of P-EMS | minutes | 27 | 24 |
| On scene time | minutes | 12 | 14 |
| Time to preferred destination | minutes | 62 | 62 |
| Patients arriving hospital alive | % | 92 | 95 |
| Debriefed responses | % | 88 | 66 |
| Adverse events | % | 3 | 1 |
| Complete documentation | % | 27 | 34 |
| Guidelines for actual medical problem | % | 15 | 41 |
| P-EMS involvement in dispatch | % | 48 | 95 |
| P-EMS necessary to provide appropriate care | % | 35 | 43 |
| Provision of advanced treatment | % | 33 | 37 |
| Significant logistical contribution | % | 50 | 52 |
| Patients enrolled in research projects | % | 0 | 10 |
| Care for relatives | % | 100 | 96 |
| Total quality score | Points (Scale: −15,15) | -1 | 1 |

Time variables are presented as medians as they are not normally distributed. The remaining QIs are presented as means of proportions.
P-EMS, physician-staffed emergency medical services.

transports and SAR responses, among others. The reason for including all kinds of P-EMS responses was to get as accurate of a picture as possible to the actual patient panorama. The reason for also including P-EMS requests without patient contact was to get an impression of safety issues, availability and P-EMS involvement in dispatch for these responses.

When interpreting quality measurements, it is important to be aware that some QI performances may intercorrelate. Imagine a mountaineer traumatised with spinal injury and neurogenic shock after suffering a fall. Packing the patient well to prevent further hypothermia and placement of an arterial line followed by vasopressors for adequate blood pressure might prevent further neurological injury—even if it takes time. In this example, too much focus on reducing on scene time could lead to a higher threshold for providing advanced treatment to correct deranged physiology. For some patients, this can be detrimental. For other patient groups, however, for example, patients with severe intra-abdominal bleeding and short transportation time to the nearest hospital, refraining from advanced treatment is likely to be beneficial. This illustrates that QIs must be interpreted with caution and that too much focus on one QI may lead to an undesired attention shift in clinical practice.

## Variability

According to Davies *et al*, there must be a certain degree of variability in the corresponding data for a QI to be meaningful.[21] If all P-EMS services report that they have 100% complete documentation every month—for example, because the electronic journal system does not allow the physicians to document incompletely—then it is not an interesting QI for quality improvement initiatives. However, a stable performance without much variation does not necessarily represent good system performance. The entire system may be uniformly underperforming, and thus goal-directed quality improvement may be indicated.

Even though the variation for a QI may be low within a single P-EMS service, there may be a high variation when assessing data from all services as a whole. When it is considered appropriate to compare single services with one another, a QI can still have enough variability to be useful. Due to the documented similarities between Nordic P-EMSs, including a comparable patient population, it is not reasonable to think that a high variability is merely a result of different case-mix.[14] It plausibly reflects real differences in performance.

## Low rate QIs

As supported by Gisvold *et al*, we conclude that events used as QIs must occur with a certain frequency.[22] In our dataset, we would describe the QI 'Adverse events' as a 'Low rate QI'. Low rate of an event limits statistical appraisal, as variation may be the result of chance. Moreover, it is difficult to use low rate of events as a continuous QI because changed rates of the event due to improvement efforts are difficult to separate from natural variation. A strictly quantitative approach to such data might therefore be less useful. However, analysing these data as 'sentinel events', where problems are studied individually to identify causal relationships and preventative measures, might be an adequate approach. Using the QI 'Adverse events' for this purpose in the future seems reasonable. When rates are too low to do statistically meaningful comparisons, qualitative data can be effective—even from small samples. Qualitative data in quality measurement can uncover issues that quantitative data may never reveal.[23]

### Documentation/validity

The validity of a QI depends on a demonstrated link between a process or a structure and a higher probability of a favourable outcome. These relationships are preferably based on scientific literature. However, where little evidence exists, these linkages can be judged important to patient outcomes by clinical experts in a consensus process.[18 24] The selection process of the QIs tested in this study is thus widely accepted.[13]

If a QI does not satisfy the criteria above (especially feasibility, rankability and variability, indicating that the variable is 'statistically' inappropriate), but the QI is still regarded clinical important, the QI may be revised to be used for the intended purpose in the future.

### Benchmarking

The data in this study are assumed representative for the P-EMS patient population and therefore transferable to other P-EMS bases in the Nordic countries. The number of responses is also relatively high. Thus, it seems reasonable to use the performances in this study as a basis for proposing benchmarks for each QI. When doing so, there are principally two approaches. The first option is to let the average score for the whole group (peer group level) serve as the average performance, and then refer to low-performance and high-performance groups related to average score. The average score will then serve as a threshold—and the aim is to perform above this level. The second option is defining a higher score, an 'excellent level' based on the performances of the best P-EMS bases. Performances above this higher level will now be the goal; in other words, this is a more ambitious form of benchmarking. How to choose the peer group is also debatable: the more homogeneous the group, the better for reliability. However, a larger group with more diversity increases the chance to learn from 'excellent performers'.[25]

According to Moore, 'benchmarking is an improvement process used to discover and incorporate best practices into an operation'.[26] When excellent performers are known, and benchmarks set, different services can measure their performance in relation to these benchmarks, which can be considered as standards. When services reach these standards, new benchmarks can be set, thus taking the quality improvement work to an even higher level. Moreover, although QIs exist for many areas in healthcare, methods to combine them into a single total score are underdeveloped.[27] We consider that the total quality score for P-EMS, as described in this paper, can be an additional tool in future quality measurement.

### Future needs

Feasible and reliable quality measurement largely depends on robust documentation systems to ensure proper data quality and to avoid added documentation workload for the clinicians. Ideally, as many variables as possible should be collected automatically through electronic data capture.

The relationship between different QI performance and a hard endpoint, such as 30-day mortality, remains unknown. Therefore, a study exploring this relationship is warranted.

### Limitations

One of the limitations of the current analysis is that the attending physicians registered all the data. They are therefore subject to registration bias and recall bias.

Except from the feasibility of the QIs, the different QI characteristics were assessed by the authors. The variability was assessed based on the data (mean and median). However, thresholds for defining poor, fair and good variability for QIs do not exist, to the best of our knowledge. Therefore, conclusions on this topic were a result of assessments and consensus among all authors. Conclusions on rankability, actionability and documentation were also resulting from assessment and consensus among the authors.

## CONCLUSIONS

In this study, a set of 15 QIs developed for P-EMS have been tested for necessary QI characteristics. The feasibility of obtaining the necessary data for these QIs was good. The variability of the QIs was adequate for all QIs except from the QI 'Adverse events', which was a 'Low rate QI'. Therefore, it seems reasonable to use this QI simply for identifying adverse events and then analyse them as 'sentinel events', rather than using these data in a quantitative analysis. The actionability was assessed poor for five QIs. Three of these QIs are measuring the timeliness of P-EMS. Some QIs depend on characteristics of the P-EMS services that might differ, such as patient volume, distances and patient characteristics; thus, they should be interpreted with caution for service comparison. However, it seems more straightforward to use these QIs for internal quality measurement of a service. To aid future quality measurements in P-EMS, benchmarks for all QIs have been proposed. In addition, we have presented a variable combining the QI performances into one single score, the total quality score.

**Author affiliations**
[1]Norwegian Air Ambulance Foundation, Oslo, Norway
[2]Department of Emergency Medicine and Pre-Hospital Services, St. Olav University Hospital, Trondheim, Norway
[3]Research and Development Unit, FinnHEMS Ltd, Vantaa, Finland
[4]Department of Anaesthesiology, Aarhus University Hospital, Aarhus N, Denmark
[5]Danish Air Ambulance, Aarhus, Denmark
[6]Airborne Intensive Care Unit, Department of Anaesthesia, Perioperative Management and Intensive Care Medicine, Uppsala University Hospital, Uppsala, Sweden

**Contribution** We thank the following physician-staffed emergency medical services for participating in the data collection: Vantaa HEMS, Turku HEMS, Tampere HEMS, Oulu HEMS, Kuopio HEMS, all Finland. Skive HEMS, Billund HEMS, Ringsted HEMS, all Denmark. Uppsala HEMS, Karlstad HEMS, both Sweden. Lørenskog HEMS, Rygge SAR, Arendal HEMS, Stavanger HEMS, Trondheim HEMS, Ørland SAR, all Norway. We thank Bjørn Henrik Moshuus, IT Manager at The Norwegian Air Ambulance Foundation, for developing the web-based database. We thank

Päivi Laukkanen-Nevala for statistical support and Jukka Tennilä for IT support at FinnHEMS Research and Development Unit, and Sasu Liuhanen at Absolute Imaginary for the adaption of the FinnHEMS database. We thank all the donors of The Norwegian Air Ambulance Foundation for the financial support that made this study possible.

**Funding** This study was funded by The Norwegian Air Ambulance Foundation.

**Competing interests** HH and AK holds research positions in The Norwegian Air Ambulance Foundation, a non-commercial charity owning The Norwegian Air Ambulance, which is the contractor of the national air ambulance service in Norway.

**Patient consent for publication** Not required.

**Ethics approval** The study was approved by the Committees for Medical and Health Research Ethics in Sweden (reference number: 2016/109) and Finland (reference number: R16031), respectively. In Denmark, application was waved by The Committee for Medical and Health Research Ethics due to the strictly descriptive nature of the study. The Norwegian Committee for Medical and Health Research Ethics defined the study to fall outside their legislation (reference number: 2016/371). This necessitated applications to The Norwegian Data Protection Authority (reference number: 16/01113-2/SBO), The Norwegian Directorate of Health (reference number: 16/14024-3) and the Data Protection Officers at the participating Norwegian health services who all approved the study.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request.



**ORCID iDs**
Helge Haugland http://orcid.org/0000-0002-5869-6675
Leif Rognås http://orcid.org/0000-0002-2542-565X

## REFERENCES

1 Porter ME. What is value in health care? *N Engl J Med Overseas Ed* 2010;363:2477–81.
2 Institute of Medicine. *Emergency medical services at a crossroads*. Washington DC: The National Academies Press, 2006.
3 World Health Organization. Quality of care: a process for making strategic choices in health systems 2006.
4 National Committee for Quality Assurance. The essential guide to health care quality. Available: https://www.ncqa.org/Portals/0/Publications/Resource%20Library/NCQA_Primer_web.pdf [Accessed 01 Mar 2016].
5. Haugland H, Uleberg O, Klepstad P, *et al*. Quality measurement in physician-staffed emergency medical services: a systematic literature review. *Int J Qual Health Care* 2019;31:2–10.
6 Snooks H, Evans A, Wells B, *et al*. What are the highest priorities for research in emergency prehospital care? *Emerg Med J* 2009;26:549–50.
7 Fevang E, Lockey D, Thompson J, *et al*. The top five research priorities in physician-provided pre-hospital critical care: a consensus report from a European research collaboration. *Scand J Trauma Resusc Emerg Med* 2011;19:57.
8 Rehn M, Krüger AJ. Quality improvement in pre-hospital critical care: increased value through research and publication. *Scand J Trauma Resusc Emerg Med* 2014;22:34.
9 Lawrence M, Olesen F. Indicators of quality in health care. *Eur J Gen Pract* 1997;3:103–8.
10 Nilsen KS, Tjelmeland KS, Halvorsen J. Olasveengen. Kvalitetsindikatorer i den akuttmedisinske kjeden. [Norwegian], 2015. Available: www.nakos.no
11 Reid BO, Rehn M, Uleberg O, *et al*. Physician-provided prehospital critical care, effect on patient physiology dynamics and on-scene time. *Eur J Emerg Med* 2018;25:114–9.
12 Institute of Medicine. *Crossing the quality chasm: a new health system for the twenty-first century*. Washington: National Academies Press, 2001.
13 Haugland H, Rehn M, Klepstad P, *et al*. Developing quality indicators for physician-staffed emergency medical services: a consensus process. *Scand J Trauma Resusc Emerg Med* 2017;25:14.
14 Krüger AJ, Skogvoll E, Castrén M, *et al*. Scandinavian pre-hospital physician-manned Emergency Medical Services--same concept across borders? *Resuscitation* 2010;81:427–33.
15 Kruger AJ, Lossius HM, Mikkelsen S, *et al*. Pre-Hospital critical care by anaesthesiologist-staffed pre-hospital services in Scandinavia: a prospective population-based study. *Acta Anaesthesiol Scand* 2013;57:1175–85.
16 Langhelle A, Lossius HM, Silfvast T, *et al*. International EMS systems: the Nordic countries. *Resuscitation* 2004;61:9–21.
17 von Elm E, Altman DG, Egger M, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg* 2014;12:1495–9.
18 Rubin HR, Pronovost P, Diette GB. Methodology matters. from a process of care to a measure: the development and testing of a quality indicator. *Int J Qual Health Care* 2001;13:489–96.
19 Eea K. Health care quality indicators project conceptual framework paper contract No: 23 2006.
20 Agency for Healthcare Research and Quality. Quality indicator measure development, implementation, maintenance, and retirement 2011.
21 Davies SM, Geppert J, McClellan M, *et al*. Refinement of the HCUP quality indicators. agency for healthcare research and quality, technical reviews No.4 2001.
22 Gisvold SE, Fasting S. How do we know that we are doing a good job - can we measure the quality of our work? *Best Pract Res Clin Anaesthesiol* 2011;25:109–22.
23 NHS Institute for Innovation and Improvement. *The good indicators guide: understanding how to use and choose indicators*, 2008.
24 Mainz J. Defining and classifying clinical indicators for quality improvement. *Int J Qual Health Care* 2003;15:523–30.
25 Romano PS. Peer group benchmarks are not appropriate for health care quality report cards. *Am Heart J* 2004;148:921–3.
26 Moore L. Measuring quality and effectiveness of prehospital EMS. *Prehosp Emerg Care* 1999;3:325–31.
27 Ken Lee KH, Matthew Austin J, Pronovost PJ. Developing a measure of value in health care. *Value in Health* 2016;19:323–5.
28 Wilde ET. Do emergency medical system response times matter for health outcomes?. *Health Econ* 2013;22:790–806.
29 Al-Shaqsi SZK. Response time as a sole performance indicator in EMS: pitfalls and solutions. *Open Access Emerg Med* 2010;2:1–6.
30 Walcher F, Weinlich M, Conrad G, *et al*. Prehospital ultrasound imaging improves management of abdominal trauma. *Br J Surg* 2006;93:238–42.
31 Brown JB, Rosengart MR, Forsythe RM, *et al*. Not all prehospital time is equal: influence of scene time on mortality. *J Trauma Acute Care Surg* 2016;81:93–100.
32 Harmsen AMK, Giannakopoulos GF, Moerbeek PR, *et al*. The influence of prehospital time on trauma patients outcome: a systematic review. *Injury* 2015;46:602–9.
33 Jauch EC, Saver JL, Adams HP, *et al*. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American heart Association/American stroke association. *Stroke* 2013;44:870–947.
34 Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. *The Lancet* 2003;361:13–20.
35 Wong MKY, Morrison LJ, Qiu F, *et al*. Trends in short- and long-term survival among out-of-hospital cardiac arrest patients alive at hospital arrival. *Circulation* 2014;130:1883–90.
36 Clark DE, Doolittle PC, Winchell RJ, *et al*. The effect of hospital care on early survival after penetrating trauma. *Inj Epidemiol* 2014;1.
37 Greenberg N. A critical review of psychological Debriefing: the management of psychological health after traumatic experiences. *J R Nav Med Serv* 2001;87:158–61.
38 Kessler DO, Cheng A, Mullan PC. Debriefing in the emergency department after clinical events: a practical guide. *Ann Emerg Med* 2015;65:690–8.
39 O'Connor RE, Slovis CM, Hunt RC, *et al*. Eliminating errors in emergency medical services: realities and recommendations. *Prehosp Emerg Care* 2002;6:107–13.
40 Institute of Medicine. To err is human. Available: http://www.iom.edu1999
41 Krüger AJ, Lockey D, Kurola J, *et al*. A consensus-based template for documenting and reporting in physician-staffed pre-hospital services. *Scand J Trauma Resusc Emerg Med* 2011;19:71.
42 Helm M, Hauke J, Schlechtriemen T, *et al*. Paper-assisted digital Mission documentation in air rescue services. Quality management in preclinical emergency medicine]. *Anaesthesist* 2007;56:877–85.

43 Gundersen L. The effect of clinical practice guidelines on variations in care. *Ann Intern Med* 2000;133:317–8.
44 NHS Centre for Reviews and Dissemination. Getting evidence into practice. *Effective Health Care* 1999;5:1–16.
45 Carnett WG. Clinical practice guidelines: a tool to improve care. *Qual Manag Health Care* 1999;8:13–21.
46 Lugtenberg M, Burgers JS, Westert GP. Effects of evidence-based clinical practice guidelines on quality of care: a systematic review. *Quality and Safety in Health Care* 2009;18:385–92.
47 Munro S, Joy M, de Coverly R, *et al*. A novel method of non-clinical dispatch is associated with a higher rate of critical helicopter emergency medical service intervention. *Scand J Trauma Resusc Emerg Med* 2018;26:84.
48 Haner A, Örninge P, Khorram-Manesh A. The role of physician–staffed ambulances: the outcome of a pilot study. *J Acute Dis* 2015;4:63–7.
49 van Schuppen H, Bierens J. Understanding the prehospital physician controversy. step 2: analysis of on-scene treatment by ambulance nurses and helicopter emergency medical service physicians. *Eur J Emerg Med* 2015;22:384–90.
50 Pakkanen T, Kämäräinen A, Huhtala H, *et al*. Physician-Staffed helicopter emergency medical service has a beneficial impact on the incidence of prehospital hypoxia and secured airways on patients with severe traumatic brain injury. *Scand J Trauma Resusc Emerg Med* 2017;25:94.
51 Bøtker MT, Bakke SA, Christensen EF. A systematic review of controlled studies: do physicians increase survival with prehospital treatment? *Scand J Trauma Resusc Emerg Med* 2009;17:12.
52 Samdal M, Haugland HH, Fjeldet C, *et al*. Static rope evacuation by helicopter emergency medical services in rescue operations in Southeast Norway. *Wilderness Environ Med* 2018;29:315–24.
53 Committee on Hospital Care and Institute for Patient- and Family-Centered Care. Patient- and family-centered care and the pediatrician's role. *Pediatrics* 2012;129:394–404.
54 Institute for Patient- and Family-Centered care. Available: www.ipfcc.org [Accessed 03 Jul 2016].
55 Dowling J, Vender J, Guilianelli S, *et al*. A model of family-centered care and satisfaction predictors: the critical care family assistance program. *Chest* 2005;128:81s–92.