Ambulance Response Programme (ARP) Impact Assessment

Executive Summary

In 2015 the Ambulance Response Programme (ARP) commenced as a component of the Urgent and Emergency Care Review under the leadership of NHS England. The programme is a clinically-led initiative set up to explore the impact and benefit of a more clinically focussed response model for ambulance providers across England. ARP is now a specific work stream within the wider Ambulance Improvement Programme, led by NHS England and NHS Improvement. The work streams within this programme include:

- Ambulance Integration Programme - This includes ARP implementation, implementation of the ambulance recommendations within the Urgent and Emergency Care Review and digital and reporting requirements
- Financial sustainability – This will focus on the identification of metrics and benchmarks, securing transferrable savings from the acute programme (corporate services, procurement, and the model hospital), clinical and workforce productivity, facilities and non-pay and collaboration with other services.
- Workforce development – This will review workforce demand and supply, workforce development, paramedic re-banding implementation, staff morale and engagement, leadership development and talent management.
- Commissioning development – This will be the area responsible for the development of a consistent model for ambulance commissioning, considering the model for accountable care organisations and devolution pilots and the engagement of the ambulance services in wider Urgent and Emergency care commissioning.
- Organisational development and configuration – This will focus on an options appraisal for organisational development and configuration in order to define the core operating model for ambulance services, creating consistency across all.

ARP aims to enhance patient outcomes, improve patient experience and reduce mortality by prioritising those with the greatest need. The programme seeks to ensure that all

patients receive an appropriate and timely clinical and transportation response where appropriate.

The School of Health and Related Research (ScHARR) at the University of Sheffield have been commissioned to provide an academic analysis evaluating the outcomes of the pilot, and provide the data baseline for assessing any future changes to national standards for ambulance providers. The impact assessment is largely based on the outputs and conclusions within this analysis.

The methodology for the academic analysis has been to complete a controlled before and after time series study, comparing changes in pilot sites to control sites over the time period October 2014 – March 2016. The pilot sites were South Western Ambulance Service (SWASFT) North East Ambulance Service (NEAS); South Central Ambulance Service (SCAS); West Midlands Ambulance Service (WMAS) and Yorkshire Ambulance Service (YAS), with the control sites being East of England Ambulance Service (EEAS); East Midlands Ambulance Service (EMAS); North West Ambulance Service (NWAS) and South East Coast Ambulance Service (SECAMB). This impact assessment will form part of the final evaluation of the programme.

The programme developed in a phased approach which commenced with the phase known as Dispatch on Disposition. This process was expected to provide appropriate call triage time to incidents and achieve a disposition before dispatch occurred. During this phase, there were developments to the 'nature of call' (NoC) questions; a series of pre-triage questions that enabled quicker dispatch and response to higher acuity calls.

Phase 2 of the Ambulance Response Programme piloted a new set of response codes with specific ambulance trusts; these codes had been clinically developed based on the principle of the patient receiving the right response, first time, according to their clinical condition. This approach has resulted in the ARP bringing together two pieces of work to achieve the overall aims. The impact of each work stream should therefore be viewed separately and brought together to understand the overall impact of the programme.

Ambulance Commissioners, through the National Ambulance Commissioners Network (NACN) have been engaged in the programme since it commenced. From the outset, commissioners requested that Quality, Performance and Financial Impact Assessments were completed for each phase of ARP. The evaluation from ScHARR is expected to demonstrate the impact on performance and quality outcomes. South Central and West Commissioning Support unit have been commissioned to assess other impacts and incorporate the analysis of the ScHARR evaluation to provide a holistic view.

The initial expectation from commissioners, in line with the aim of the ARP Development Group, was that the widespread adoption of ARP would deliver improved outcomes for patients. It also became clear, during the early stages of the programme, that operational efficiencies could be achieved to support improved performance against national performance standards. This assumption was underpinned by discussion at the Public Accounts Committee indicating that ARP should produce financial efficiency through operational efficiency gains. In line with the National Urgent and Emergency Care review it is also expected that material changes should offer wider system level performance or financial utilisation benefits.

As discussed, the output of the ScHARR report has been used to demonstrate the impact of ARP on performance and quality. The financial impact was not within the scope of the academic study. SCW has therefore undertaken this piece of work through direct engagement with the trust pilot sites which identified significant challenges and resulted in a less financially objective assessment. The obstacles to achieving a purely objective financial assessment resulted from the variation in starting position of each trust, with the variations evident in pre and post ARP implementation baselines in the following areas; performance, fleet ratios, demand, workforce, geography, current funding arrangements and the required capital investment to deliver a new operational model.

During the ARP pilot there has been a high level of scrutiny in the ambulance service, including a National Audit Office (NAO) report and a review by the Public Accounts Committee (PAC), both of which included recommendations which is anticipated can be managed through ARP and the Urgent and Emergency Care Review. There has also been a review in pay grade of the paramedic workforce which has resulted in a banding increase. This workforce change is relevant on the basis that the role of the ambulance service has evolved and the autonomy of the Paramedic role increased. This has led to a shift in the expectations which are placed on both to enable better outcomes, such as improved clinical decision making supporting the delivery of care in the most appropriate environment, and ultimately supports the aims of ARP.

The recent NHS Ambulance Services National Audit Office (2017) report [1] stated that commissioning of ambulance services was not consistent across trusts. There were several recommendations which came out of this report, two of which commissioners are expecting the outcome of the ARP to support to some degree, although it recognised that full delivery is likely to be through the wider Ambulance Improvement Programme. These recommendations are as follows

- NHS England, NHS Improvement and ambulance trusts in England should work together to define the optimal operating framework for an ambulance trust, allowing some flexibility to tailor responses in urban and rural areas.

- Ambulance commissioners should take a consistent approach to commissioning ambulance services, based on the framework. As part of a standard operating framework, trusts should develop and report consistent metrics on efficiency, including staff utilisation (within the report the indication is that lack consistency in the funding arrangement and pricing for ambulance service contracts is a key contributor to this issue).

The recent Public Accounts Committee report [2] also makes some recommendations some of which it is anticipated should be delivered through the ARP as follows:

- The Department of Health, NHS England, NHS Improvement and ambulance trusts should implement the recommendations of the Ambulance Response Programme at pace. Any changes to the response-time target system should address 'tail breaches' (very long delays) and the lack of focus on Green calls.

- The Department of Health, NHS England and NHS Improvement should set out a trajectory with clear milestones for all its modernisation programmes that focus on ambulance services, by October 2017. As part of these programmes, they should ensure consistent and reliable data sets for key performance measures are available, including clinical outcomes, new models of care, efficiency metrics, and patient-transfer times at hospital.

- NHS England and NHS Improvement should assess whether sufficient resources are available to ambulance trusts to support new ways of working including capital expenditure.

Delivery of the Urgent and Emergency Care Review outcomes is expected to be achieved through the Urgent and Emergency Care Networks (UECN) and the oversight boards delivering Sustainability and Transformation Plans (STPs). STP Boards expect this to be delivered through commissioner and provider collaboration that achieve overall system benefits. Ambulance specific commissioning has historically been a discreet area of commissioning, however in recent years the performance and financial impact of ambulance services have increased within local systems. Subsequently, health system partners have increased their understanding of the need to engage with the ambulance service as a key player within local systems and subject matter expertise in commissioning ambulance services has increased.

The Ambulance Response Programme is viewed by ambulance commissioners as a potential enabler to support wider system changes in the development of new models of care. In line with the Urgent and Emergency Care Review, the ambition across STPs is for more patients to be treated in the most appropriate setting, with fewer patients conveyed to emergency departments where this is not clinically indicated. Initiatives across STPs will need to provide access to alternative care pathways, or access to multi-disciplinary clinical advice, to support ambulance service clinicians to deliver appropriate care in the community. Therefore it is essential that the benefits of ARP are considered in the context of the wider system. This will require additional support to ensure STPs are fully sighted on the potential benefits of ARP and in planning for the integration of ambulance services into the STP urgent and emergency care delivery plans.

In summary the impact assessment has found that phase 1 of ARP maintained stable response time performance for the highest acuity patient cohort (Red 1), however there was a significant improvement for most trusts in the Red 2 cohort which are also identified as high acuity. Phase 2 is more challenging to assess because of the significant model change. However ARP appears to have enabled stable response time performance for all categories during a period of winter pressure with associated high demand. As a comparison, non-pilot site trusts showed deterioration in current model performance over the same period. Assumptions can therefore be drawn that phase 2 also has a positive impact on performance

Clinical outcomes results are inconclusive due to the volumes involved and the time period over which they have been assessed. However there are early indications that stability has been brought to previously deteriorating measures. Further monitoring will be required in order to fully understand if ARP enables the current trend to be reversed.

It is very clear that ARP creates operational efficiency. The leading example of this is the information from West Midlands Ambulance Service. The impacts of these efficiencies vary between trusts. This is due to the number of variable factors which influence this outcome, mainly the baseline position of the trust in relation to pre and post ARP implementation performance, fleet and workforce ratios. For trusts with favourable baselines the operational efficiency is significant and allows consideration of resource streamlining. For other trusts this is not case and ARP becomes an enabler to improving the position against these factors.

The operational efficiencies evidenced allow an assumption that financial benefits can be realised. The evidence points to this being cost avoidance rather than cash releasing, and at best allows trust to invest in improved quality outcomes. Actual analysis of the financial

benefit has been a challenge. Good engagement with the ambulance trusts has allowed a narrative approach to be taken which shows positive results.

There are impacts which need to be considered as offsetting some of the benefits under ARP. These include fleet and work force. ARP may require a significantly different operating model including revisions to fleet and work force ratios. This requires potential capital investment for the fleet and potential ongoing work force increases with associated costs. The impact is again variable by trust due to the baseline positions. Each trust, in conjunction with their commissioners, will need to fully consider these two factors to understand if the operational benefits offset the cost implications.

In summary there is strong evidence to support the case for change to ARP. There are some risks, particularly in potential work force and fleet investment requirement which need to be considered at a local level.

It is clear that the current model is no longer fit for purpose, has driven inefficient behaviours and requires review. ARP has shown enough evidence that, in comparison to the current model it is likely to be an improved solution, particularly if there is scope to maintain a review and improve approach to any agreed roll out.

Aim

The aim of the ARP has been to improve patient outcomes and increase the operational efficiency of ambulance service provision. As the programme evolved, it was also suggested that ARP may generate financial efficiencies across the ambulance service, with the potential for more to be achieved at a system level. This impact assessment examines these assumptions and seeks to quantify the impact on the ambulance sector and indicate opportunities for efficiencies in the wider system.

The assessment intends to follow the format of the ScHARR report by making an assessment on Dispatch on Disposition (DoD) (as phase 1) and then of the clinical code changes (as phase 2), bringing them together in order to assess overall impact, including financial.

This impact assessment will also look to demonstrate the potential for further development, or the link to other known pieces of work, with particular focus on relevant recommendations from the National Audit Office (NAO) report and the Public Accounts Committee (PAC): potential implications for workforce and any opportunities for the wider system from increased levels of Hear & Treat (H&T) and See & Treat (S&T).

Some of the potential benefits of ARP remain dependent on the implementation of further system wide changes to support the aims of ARP. Commissioners will be required to take

this into consideration through the local STP work streams, using the output from this and other linked work streams as the basis for development. It is also recognised that realisation of all potential benefits under ARP will take time, due to the significant cultural and operational changes required, both within the ambulance sector and in the system response.

Impact Assessment

The majority of the impact assessment is based on the data and analysis within the ScHARR report for phase one and two of the ARP which assessed the effects of DoD and the coding changes. The impact of each phase will be looked at separately and then brought together for an overall view.

Throughout the trial, information has been provided to ScHARR on 30 different indicators (Appendix 1) which have been measured pre-and post-implementation of each phase. This impact assessment will group these together in the following domains where possible:

- Performance – pre-and post DoD against the current national standards of Red 1 and Red 2.
- Clinical outcomes
- Operational Efficiencies – an assessment of the indicators which show benefits to allocation and appropriate use of ambulance resources.
- Fleet and Workforce.
- Incident outcomes – an assessment of the impact of any changes to clinical outcomes and also Hear and Treat or See and Treat rates.

The impact assessment will also take into consideration the impact of ARP on the following

- Wider system benefits including NHS 111 impact
- Financial Impact

Performance

Phase 1 - Dispatch on Disposition (DoD)

According to the ScHARR report there was no statistical difference between the pilot sites and the control sites in four of the performance indicators. Two of these indicators related to time of response in relation to Red 1 category calls - this indicates that DoD does not improve the response time to the higher acuity calls against the NHS Constitution requirements compared to the previous approach, however neither does it show deterioration.

As the impact for this patient cohort is shown to be neutral it has been assumed that there is no increased patient safety or outcome risk associated with DoD, however, it is of note that tracking of patients for clinical outcomes throughout their care pathway was not incorporated into the ARP evaluation process. This would be a significant piece of work, currently outside of the scope of ARP, made challenging through the requirement for full system engagement and data sharing agreements. However it could be possible for local systems to test this through triangulation of existing data feeds.

Patient experience within this group is assumed not to be affected due to the neutral effect on Red 1 performance, on the basis that no serious incident was reported by the ambulance services within the pilot.

For the Red 2 response, the data within the ScHARR study showed a performance target improvement of 5.8% when reviewing all trusts (model 2). This is a significant improvement, following a period of time where trusts have struggled to improve response targets under the previous model. There is patient benefit to this as this remains a high acuity patient cohort gaining an improvement in time to response by the ambulance service. However, as the evaluation did not track the whole patient pathway it cannot be evidenced that this improvement led to improved outcomes for patients on discharge from either the ambulance service or the receiving provider.

During the initial trial phase of DoD it was noted by SWASFT, as the early implementer, that there was an unintended consequence of a reduction in Red 2 performance, This was due to a reduction in the percentage of calls with a DX014 ‘early exit incident’ disposition, all of which are categorised at Red 2. This occurs as dispatch is delayed until the triage process is completed in the call centre and subsequently, there are significantly fewer incidents where the ambulance service resource arrives on scene prior to triage completion. On further roll out, the data shows that there is a general downward trend for most services in the percentage of Red incidents which are coded DX014. The drop in the percentage of DX014

was an immediate effect of DoD introduction for each trust and very quickly stabilised. This produces a 'technical loss of performance'. It does not necessarily equate to a lower level of response for higher acuity patients as more patients are being appropriately triaged with associated response codes. The impact of this is to show true performance against actual patient triage outcomes, as correct dispositions are reached more often and should therefore be viewed as positive. Despite this technical loss, overall performance has increased for Red 2 as discussed above.

There were three measures where the control sites showed 'better performance' than the pilot sites. These were median time to resource allocation, and median and 95th percentile time from call connect, to resource on scene for Green 2 incidents. All of these relate to Green 2 calls which, though classified as lower acuity calls, include patients who have a clinical condition requiring a timely response e.g. drug overdose or potential fractured hip dispositions.

There is a degree of clinical risk associated with this patient cohort, assumed to be lower than Red 1 and Red 2. The acceptance of this risk has been a core element of the ambulance triage process for over a decade; however, commissioners and providers have come under increasing scrutiny of this when excessive delays in responding have occurred. In developing new models of urgent and emergency care at an STP level commissioners and providers may need to consider clinical variation and risk at a population and system level rather than by individual provider. The increasing concern around performance is the achievement of reasonable response times to the 'tail' of these incidents, i.e. those with the longest waits. The data within the ScHARR report indicates that long waits are further extended despite the introduction of DoD. This remains a concern and it is clear that the current performance regime does not encourage ambulance trusts to give focus to these. This concern is supported by the recommendations from the PAC and new metrics which require a more holistic view of incident response would address this. The overall ARP has therefore sought to develop these new metrics, in order to increase the focus on longer waits. It is anticipated that these new metrics will be rolled out with the full ARP programme.

<u>Phase 2 – coding trial</u>

Performance comparisons under phase 2 of ARP become more challenging. This is because phase 2 represents such a significant change, requiring completely different approaches to response that pre and post implementation performance cannot be compared directly. The code changes significantly reduced the proportion of incidents requiring an eight minute response. The ScHARR report manages this through the use of a descriptive analysis. However the focus of ARP has been to ensure that higher acuity patients are provided with a timely response and, whilst not equivalent, the performance trends are shown across all

three phases (DoD, code change 2.1 and code change 2.2) to allow some discussion. These show there were some differences between the three pilot sites. One site (WMAS) showed steady performance response times with the expected effect of increased fluctuations in demand leading to decreased performance.

Post implementation of Phase 2 however, WMAS showed a decrease in performance not associated with increases in demand, which levels over time as the change becomes embedded. In another trust (SWASFT) the opposite occurs and there is an increase in the proportion of the most urgent calls responded to within eight minutes relative to demand. In the third trust (YAS) performance remains stable, although the effect of demand beyond a certain point can be seen in terms of widening the gap between demand and performance. The reducing performance in WMAS is explained as the trust implemented a number of operational changes which were required to support the implementation of ARP coding change.

In summary, there are no statistical change trends in the majority of measures under ARP 2.2. However this was achieved during a period of winter which included demand peaks and increased impact seen in handover delays at acute trusts. This suggests that the new operating model has supported the maintenance of performance when trusts have been under significant pressure and should therefore be viewed as a positive impact on performance, with the focus being on the delivery of faster response times to those with the most clinical need. The ScHARR report states that the trial has only been operating for a relatively short length of time and clearer changes may become more apparent if this new model of service is delivered and evaluated over a longer period of time.

Urban versus Rural performance

There are long standing assumptions across the sector that delivery of rural performance is more challenging than in urban areas due to geography, the dispersed population and access to secondary providers. This is also recognised in the commissioning of ambulance services as current contractual performance and quality standards are set at averaged trust levels and an understanding that there may be cost implications for rural areas to improve these standards.

The impact of ARP on equity of access across rural and urban geographies has therefore been measured, with complex results. They show that for some measures ARP has reduced the impact of urban versus rural geographies on performance. However some of the analysis challenges the initial assumption, in that for a substantial number of measures, response times were longer in predominantly urban areas than in mixed or rural areas in all three pilot sites, particularly in relation to the 95th percentile times.

There are clearly several factors which need to be considered here, such as operating model, actual population density and associated demand, traffic in urban areas but shorter travel distances and availability of resource. This requires more work to fully understand, but is becoming increasingly relevant as STPs look to develop their system responses to meet the needs of their population.

Overall this performance analysis demonstrates the original aim of ARP, in terms of ensuring the sickest patients get the fastest response, has been delivered, particularly with the performance increase witnessed with DoD nationally. The standards which have been used to measure performance within phase 2 also offer the right environment to enable the development of the optimal operating framework for ambulance services as recommended in the NAO report which must take place before the commissioner and provider discussions can progress in terms of aligning commissioning intentions with the operating framework, which is the second recommendation.

Clinical Outcomes

One of the objectives of ARP is to improve clinical outcomes for patients. ScHARR have used the three Ambulance Clinical Quality Indicators (ACQIs) of Stroke (time of call to arrival at Hyper Acute Stroke Centre), STEMI (time of call to Primary Percutaneous Coronary Intervention) as a measure of clinical outcomes and cardiac arrest (return of spontaneous circulation and survival). Of these, only the cardiac arrest indicators have a genuine outcome measure, rather than just an intervention, however the other two are considered as best practice. ScHARR cite challenges in assessing the impact of ARP on these due to the small volumes of case numbers and baseline variations in month on month performance so the impact of ARP versus other factors are difficult to detect.

The time lag in publication of data also presents a challenge; this is because of the need to triangulate data with the Myocardial Ischaemia National Audit Project (MINAP) and Stroke Improvement National Audit Programme (SINAP) data sets. However the limited information available to ScHARR does show stable performance for most indicators apart from Stroke outcome which shows a downwards trend. The latest NHS England figures show that this downward trend is stabilised in all three pilot sites with improvements into December for two sites and stable performance in the other. This compares to a national picture which continues to deteriorate. Even with this updated information, which then covers the second code set change, it is not yet possible to determine if this is a trend which can be sustained through Phase 2.

When reviewing this it is essential that STPs consider any changes in HASU or PPCI provision which will have an impact on ambulance conveyance times and performance.

Early detection of cardiac arrest in the call taking cycle has shown an improvement within ARP through the introduction of the Nature of Call (NoC) which is 3 pre-triage questions and nature of call identification using a pre-defined list of problems. The ScHARR report states that across all sites there has been around a 70% capture rate of cardiac arrest through the use of NoC which enables a faster overall response for this patient cohort. This only represents around 0.6% of total incident volume, but is the cohort for which speed of response and subsequent defibrillation is the key action to improve the outcome and so this is a positive effect.

Operational Efficiencies

Front Line Operations (response)

The study used twelve indicators to measure allocation of resource changes under ARP with the intention that these would be an indicator of efficiency (Appendix 1). The result from the implementation of DoD showed a consistent pattern of a reduction in resource use across all twelve indicators. The measures used in these indicators have allowed ScHARR to make a national estimate on resource gain.

ScHARR estimate that there is potential to gain 10,243 whole resources which would be available at the time of 999 calls to respond per week on a national basis, through DoD. The same efficiency measure has been reviewed for the phase 2 pilot with the ScHARR estimate suggesting this is 5,697 further resources, which gives a total of 15,940 resources.

However the definition of this measure needs to be clearly understood before any further assumptions can be made. The estimates have been derived from the cumulative effects of reducing the resource per incident, through reduced double or triple dispatches and reduced stand downs, so that eventually an additional vehicle is available for response; these wholes are made up of fractions in each case. It indicates availability at the time of incoming calls but does not give a measure which can be converted into something more measurable in terms of efficiency such as unit hours. However this is still a significant operational efficiency that providers and commissioners would want to appropriately realise.

There is a description within the ScHARR report, which has been provided by WMAS, that gives some strong evidence of the operational efficiencies they have experienced. They have been able to reduce the average response per incident from 1.3 to 1.1 and have demonstrated a requirement for 4% less overall resource to deal with an additional 10% of demand through the change in the operating model.

Clinical Hub (call taking, dispatch and remote clinical decision making)

ARP will not impact on the total number of calls or incidents which the ambulance services receive; therefore there is no possibility of realising any operational efficiency gain at the initial point of patient contact. If the volume of calls remains the same or increases as is the trend, then the volume of staff required to field these calls will need to increase proportionally to this.

Call taking resource requirements are driven by the individual calls and the clinical decision support software (CDSS) used. There are two approved CDSS which trusts are able to use; Advanced Medical Priority Dispatch System (AMPDS) or NHS Pathways. Of the pilot sites YAS uses AMPDS, WMAS uses NHS Pathways and, due to legacy issues from trust mergers, SWASFT currently uses both for different areas of the trust. The different systems require differing levels of each resource as call takers using NHS Pathways can Hear and Treat more calls, but AMPDS requires more clinicians to complete the triage. The ARP development and delivery groups have been very clear that assessing the most effective CDSS under ARP has not been within the scope of the programme. These discussions and decisions would need to be evaluated at a local level.

Using SWASFT as the example site, they have reported increased Hear and Treat rates compared to historic levels; however the opportunity to deliver these increases can be attributed to both DoD and the coding changes. It is not possible, at this time, to identify the improvements attributable to either specific development.

SWASFT has invested in more clinicians to support Hear and Treat but the true impact of this has been limited by the ability of the trust to recruit suitably qualified and experienced staff. SWASFT assume that a greater impact could be achieved under ARP should the optimum level of clinician in the control room be achieved.

There is potential for this to be achieved through delivery of other strategic commissioner objectives within the Urgent and Emergency Care review, such as the implementation of Integrated Urgent Care hubs. This would allow more calls to be transferred to a service which may enable increased hear and treat outcomes for patients and allow them to manage their healthcare needs in the most appropriate setting, without the need of a face to face ambulance review.

SWASFT make the assumption that any increases in hear and treat should result in a reduction in the number of see and treat incidents, and will enable both a reduction in the number of resources that are stood down and a reduction in multiple dispatches to incidents creating further operational efficiency gains.

Wider System Benefits

There is the potential for this to become a significant benefit to the healthcare system as increased resource availability should allow ambulance services to better respond to all incidents. There are significant issues created through delayed response to ambulance incidents, for healthcare systems; of note are delays in the conveyance of Health Care Professional booked calls where, notwithstanding the immediate clinical risk to the individual patient at the time of delay, it is widely recognised that later conveyance to the Emergency Department (ED) can add significant delays to the patient assessment, increase likelihood of admission and negative impact on other patients accessing the system.

The response delays can result in ambulance to ED handover delays, increased time spent within the ED, reduced ED performance against the 4-hour standard, increased potential for admission in the out of hours' period as patients are presented later in the day, increased overall length of stay and potentially increased morbidity. If the operational efficiency was fully realised it could have a positive impact in supporting patient flow or enable resource to be released to increase capacity elsewhere in the local healthcare system.

Fleet

Discussions have been ongoing to support a comprehensive impact assessment of ARP on ambulance service fleet. Due to the number of variables involved, a consistent message on the impact creates a significant challenge to describe. However, the early indication is that for ARP 2.2 to support improved performance, and to dispatch the most appropriate resource first time, a different fleet arrangement may be required than under previous conditions. Previously some ambulance providers invested heavily in Solo Rapid Response Vehicles (RRV) to support achievement of the 8-minute standard with a Double Crew Ambulance (DCA) being used to transport patients to hospital. Under ARP it appears more appropriate to have a configuration more weighted towards the DCA.

The three ambulance providers involved in the trial of the new response codes each have a different fleet configuration, and the extent to which fleet changes need to be managed are still not clear. They are likely to differ between trusts based on various factors including rurality, activity, mix of category and changes in operational delivery models. Specific consideration needs to be given to achieving the optimum ratio of conveying resource (DCA) against RRV to support delivery of a clinically focussed response model.

The ARP development and delivery groups have discussed national level modelling of this, however, due to the variables described above, it was agreed that delivering national modelling would not give enough granularity at a local level to achieve clarity for each

individual trust and that this should be managed locally. The variables between trusts in the baseline ratios in current model fleet requirement versus ARP means that there will be a need to review the capital investment and potential work force increases at a regional level.

Workforce

Though ARP is described as a clinically focussed response model for ambulance providers, it should also be recognised as one of the most significant change management initiatives undertaken across the ambulance service for a number of years. The formal report from ScHARR provides analysis from the operational staff who have taken part in a number of staff surveys, to gauge their reaction to ARP. This has largely been positive with most staff viewing ARP as positive.

ARP requires a different model to deliver. This has an impact on workforce as trusts look to decrease the ratio of solo responders and increase the ratio of double crewed resources, so the workforce requirements increase. This may be in the form of non-clinician support to the paramedic role, but this needs to be understood as a potential risk in associated costs, recruitment and retention.

Alongside the cultural changes required, there will need to be an assessment of the number of clinicians and the level of clinical skills required to support changes in care delivery, such as increasing Hear & Treat and See & Treat outcomes. Delivering increased Hear & Treat may require an increased investment in the number of clinicians within the clinical hubs of ambulance service, or integrated urgent care services. Delivering increased See & Treat is interdependent with the competency skill base of the newly agreed band 6 paramedics.

This will require alignment with the workforce development programme within the Ambulance Improvement Programme to ensure that the band 6 paramedics possess the clinical skills required to support this model of care. This needs to be aligned by the STP Boards developing their workforce plans to increase community based care, including consideration of the potential of ARP.

Retention and recruitment of paramedics is recognised as a significant challenge for ambulance providers, with the associated financial impact both in the time spent recruiting and training new members of staff. A significant financial impact is the need to utilise premium rate third party resources to achieve daily establishment. The operational efficiencies evidenced by ARP could help to mitigate this issue, as staff satisfaction increases and resource requirements are reduced.

To further benefit, Ambulance Trusts and STPs will need to consider how the workforce can be reshaped to allow development of the ambulance workforce to realise the aims of ARP in

sending the most appropriate resource based on the clinical need of the patient, making better use of the whole health care workforce across the ambulance sector and the system.

The issues relating to ambulance workforce are considerably broader than ARP, and include the recent move to band 6 paramedics and up-skilling of the workforce to deliver additional hear and treat and see and treat. This is being picked up through the workforce work stream within the NHS improvement led Ambulance Improvement Programme, but is worthy of note here to understand the link with ARP as an enabler to improvements.

NHS 111

ARP coding changes have the potential to support further system wide benefits within urgent care. Specifically this relates to NHS 111 ambulance dispositions. Since the implementation of ARP coding changes, it has been noted by the ambulance service that the proportion of incidents flowing from NHS 111 requiring an immediate response is significantly less than under the previous categorisations. Using SWASFT as the example, it can be seen that the proportion of category 1 calls which come through from NHS 111 are approximately 1% of the total volume.

Under previous categorisation, immediate response incidents with no opportunity to re-triage would have consisted of all Red 1 and Red 2 calls, making up around 40-50% of the ambulance dispositions produced by NHS 111 (varies between ambulance and non-ambulance providers). SWASFT report that this has enabled significant operational efficiencies to send the most appropriate resource under ARP and, increased the potential to achieve a hear and treat outcome for the lower category calls. There is consistency in this across both ambulance and non-ambulance providers of NHS 111 services.

Currently there has been no move within ARP to achieve the conversion of NHS 111 dispositions to the ARP 2.2 code set at the point of source and the process remains within the CAD of the pilot sites. However, if ARP 2.2 was incorporated into NHS111 dispositions there is a significant opportunity to achieve a hear and treat disposition within the NHS 111 environment, or an integrated urgent care hub. This should enable enhanced patient experience with reduced handovers between providers, and increased potential to manage the patient in the most appropriate setting by the most appropriate service.

ARP therefore needs to be taken into consideration as commissioners look to procure new services of NHS 111 under the new models of integrated urgent care, ensuring that the appropriate resource is available to support improved outcomes from NHS 111. This will need engagement from the ambulance service through the STP Board and lead commissioner locally.

Financial Impact

Due to current financial constraints commissioners are requiring a robust Quality Adjusted Life Year or Quality Impact Assessment of all investment / decommissioning decisions, which has proven historically challenging with ambulance services. The historic lack of robust financial data means that commissioners have not been able to consider ambulance service commissioning on equal terms to other services they are commissioning within the urgent and emergency care pathway, or measure the impact of a more effective pre-hospital ambulance service to the end patient outcome.

Understanding the financial impact of ARP has also been a challenge. Again the position nationally presents so many variables in terms of individual trusts that a national position statement is all but impossible. The intention had been to take the metrics being produced within the ScHARR report and convert this into something measurable by the ambulance trusts to produce financial efficiency figures. However the measures within the ScHARR report are not translatable into such a measure (for example unit hours) they are an indication considered at a single moment in time.

There are several considerations which must be made to understand the financial impact of ARP. Firstly, most ambulance services start from a place where they are not meeting current operational standards. DoD improves this for Red 2 response but may not bridge the gap to achievement. The only solution for further improvement, unless further change is considered, is therefore increased resource targeted to these incidents. How this investment can be achieved is dependent on commissioner positions and/or the ability to realise other efficiencies through ARP, or within the Urgent and Emergency Care system. For the trusts, the financial implications are similar, in that internal and external investment options must be considered if improvements are to be achieved. Current positions and trust baselines must be considered when reviewing these options For example a performing trust without resource against demand shortfalls would be able to make decisions around investment in quality improvement, but trusts that are not in this position would need to take into account the need to manage demand, increase performance and quality improvement together.

The release of capacity already commissioned through the operational efficiencies identified in ARP indicates that investment requirements to achieve NHS Constitution standards are reduced, or can be targeted at more specific improvement requirements, such as the delivery of Category 1 performance, the management of delayed responses or increases in demand. The trust would also potentially see a reduced financial risk associated with contractual penalties against non-achievement of targets.

It is however also clear that the implementation of ARP can have a negative financial impact, at least in the short term. It has been discussed here that fleet ratios may need to be considered and changed, in some trusts significantly, to meet the operational delivery model requirements. This requires investment, firstly in a capital expenditure, but also in the increased associated costs of the different fleet, e.g. depreciation, fuel and significantly staff. This will offset at least some of the financial efficiency within the operating model.

More detail on individual trusts has been difficult to achieve given the variables and so through engagement with trusts SCW has secured specific examples, given below, of how trusts have used the operational efficiencies in the decision making on improvements, be this performance, quality or workforce requirements. What has been clear through this process is that ARP is an enabler to cost avoidance rather than achieving any kind of cost saving. The extent of this cost avoidance is different per trust.

Trust Examples of the Benefits of ARP

The following statements have been developed by the pilot site trusts following engagement with SCW and have been added without SCW edit.

YAS Examples of Benefit of ARP

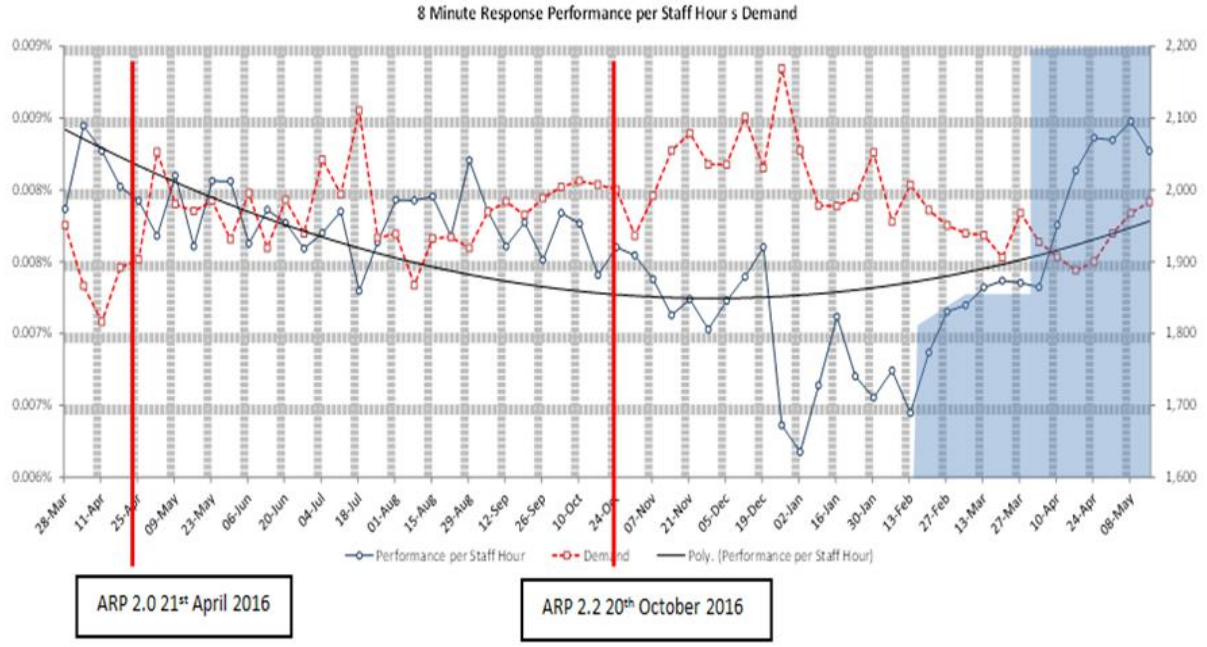
In a steady state a reduction in vehicles per incident increases the availability of crews to respond to incidents and therefore generates a performance improvement. Alternatively it may be possible to reduce resource whilst maintaining the available hours, which in turn will maintain performance. However, this is predicated on having enough instances of availability in a location to be able to remove a responding vehicle from the rota.

With the increased focus on transport resources in ARP there is a marked shift in the vehicle ratio required, i.e. more double crewed ambulances (DCAs) and less rapid response vehicles (RRVs). This impacts both the number and skill mix of staff required. Additionally, the requirement for more DCAs has increased capital and revenue expenditure due to the vehicle and equipment purchase as well as the ongoing revenue impact from increased deprecation, PDC, higher maintenance costs and reduced fuel efficiency. However, we currently are not at our optimum mix of fleet to deliver the overall benefits.

In terms of YAS specifically, due to increasing demand during the pilot and the challenge to meet performance targets the efficiency generated has part offset the increased demand, part offset increased handover time and part offset performance challenges. It has therefore been a cost avoidance measure rather than a cost saving.

For YAS, the current 999 contract is based on demand and performance projections under the ARP delivery model in terms of staffing. The cost avoidance efficiencies are therefore already crystallised. The Trust is, however, challenged in terms of meeting the fleet requirements of the combination of demand increase and change in vehicle ratio both in terms of capital and revenue requirements.

A point to consider is that the evidence from YAS suggests that if rotas and vehicle mix are not updated to reflect the change in focus brought about by ARP it could cause a reduction in efficiency and performance. The graph below shows that performance per staff hour, a measure of efficiency and performance, falls significantly post ARP 2.2 until the rotas and vehicle mix are changed (blue shading). However, given this reduction covers the challenging winter period it is not possible to say with certainty how much of the impact is due to winter demand, ARP or better alignment of rotas.



SWASFT Examples of Benefit of ARP

SWASFT does not have a Patient Level Information and Costing System (PLICS) and therefore is not able to provide the level of costing data requested initially. The response therefore fulfils the alternative agreed approach focused on narrative and with an emphasis on hours.

The cost implications needs to be clear if this is just the impact of ARP 2.2 (change in clinical coding) or whether this should include the impact of the Dispatch on Disposition trial (increased time before dispatch of resources).

ARP 2.2 does not impact on the total activity received by the Trust and does not impact on call taking resources which are driven by the individual calls and triage system used.

SWASFT currently uses both AMPDS and NHS Pathways although a decision has been made on which common system to use. The different systems require differing levels of each resource within the hub as Call takers using NHS Pathways are able to Hear and Treat more calls but AMPDS requires more clinicians to Triage.

SWASFT have reported increases in Hear & Treat rates compared to historic levels, however the opportunity to deliver these increased levels can be attributed to both DOD and ARP 2.2 and it is not possible at this time to identify the improvements attributable to either development.

SWASFT has invested in more Clinicians to support Hear and Treat but the true impact of this has been limited by the ability to recruit suitably qualified staff.

Any increase hear and treat should result in a reduction in the number of see and treat incidents and will enable both a reduction in the number of mobilisations that are stood down and a reduction in the multiple resources mobilised to an incident. This is based on sending the correct resource first time.

This has a productivity benefit for frontline resources which has been estimated as follows:

Reduced Stand Downs

- Had the Trust mobilised its available resources in similar priorities/processes used pre-Arp the Trust would have reported in the region of 13,000 additional stand downs compared to the actual number reported between 25 October and 31 December 2016. In reality the reduction that can be directly attributed to ARP is likely to be lower as the number of allocations is likely to have reduced due to other operational factors. If we assume that just 50% of this improvement relates to ARP, 6,500 fewer stand downs across a 68 day period equates to around 35,000 fewer

stand downs on an annualised basis. Based on 5 minutes resource time saved per stand down this equates to a saving in the region of 2,900 resource hours per annum.

Reduction in Arrivals at Scene

- By sending the most appropriate resource to scene the Trust has reduced the number of resources per incident, avoiding unnecessary duplication of resource time. Across the 68 day period, when compared to the response profiles in Q4 of 2015/16, the Trust utilised the equivalent of 2,680 fewer resource hours on job cycles. Again other factors will have influenced the change in resource hours including demand, conveyance rates, seasonal pressures and mix of workload. However annualising these hours equates to 14,390 hours per annum and if we again assume 50% of this reduction can be attributed to ARP then this equates to 7,195 resource hours saving per annum.

- The combined effect of reduced resource utilisation adds up to around 10,000 resource hours per annum.

The ARP 2.2 should not impact on conveyance rates as this is based on a clinical decision rather than the change in coding within the hub.

The combined impact of DOD and ARP2.2 should be an enabler to review the resource requirements of the organisation but there is a lead time for this which has yet to be fully implemented and tested.

The change in the operational fleet mix increases the number of conveyance resources (Dual Crew Ambulance, DCA) and reduce the number of Rapid Response Resources (RRV) single paramedic crew.

This change requires an upfront capital investment in the more expensive fleet i.e. DCAs cost circa £130k compared to an RRV circa £35k alongside associated medical equipment and the additional staff costs alongside the Paramedic. The cost per unit hour of resource increases by circa 50% for each hour changed. The level of change in each area will be dependent on the local geography, current resource mix and the funding available.

It is not possible at this stage to assess whether the benefit in reduced stand downs and reduced arrival at scene offsets the additional costs associated with the capital investments.

It should be noted also that the initial benefits described in the reduced stand downs and arrivals on scene have already been taken by the Trust in managing demand over 2015/17

and that these benefits have been constrained by the existing level of resource which is based on the "old" operating framework.

SWASFT secured additional activity funding as part of the 2016/17 999 contract. Included in his figure was £3.6m for the investment in 16 Dual Crew Ambulances across the Trust. The Trust operates within the ARP trial and commissioned a review of its operational rotas from ORH taking into account the revised targets. The Trust has made the decision to change its rotas across the region, one of the drivers to this was a response to ARP but also to increase productivity. The revised rotas implement a rota model with an increased % of Dual Crew Ambulances rather than Rapid Response Vehicles. As part of this process the Trust has utilised the additional resources and the core funded resources to create a new rotas plan to be implemented in Q2 of 2017/18. This rota has converted circa 60 RRVs to become DCAs with the existing resource hours. The capital cost of a DCA is circa £130k compared to an RRV of £35k this has required an additional investment of £5.7m capital to implement the rotas following the implementation of ARP.

WEST MIDLANDS AMBULANCE SERVICE NHS FOUNDATION TRUST

June 2017

Introduction

The Trust has been asked to provide a narrative description of the operational benefits which WMAS has witness and been able to implement whilst working under ARP 2.2 methods.
WMAS went live with ARP 2.1 in early June 2016, and latterly went live on revised ARP 2.2 in October 2016.

What has been done and achieved?

One of the key changes which has been achieved under ARP2, is to reduce the number of RRV resources and increase the number of on duty Emergency Paramedic Ambulances. This has the key benefit of ensuring the Service has adequate transportation capacity, instead of workload building up each day, RRVs experiencing delayed backups and patients experiencing delays. It also eradicates the incidents of Paramedics on RRVs needing to travel with a non-Paramedic Ambulance for onward patient care. Before ARP was implemented the Service would typically provide the following resource at peak of day, 99 RRVs and 215 Emergency Ambulances (snap shot taken from 1 July 2015), verses now

providing 14 RRVs and 310 Emergency Ambulances.
The above changes, which have been implemented to optimise the operational arrangements whilst operating under ARP ver2.2, have provided the following benefits:

Quality

The Trust has been able to improve the quality and consistency of Service whilst at the same time improving efficiency. The table below sets out a number of tangible and measurable benefits of operating differently, which has been facilitated by the implementation of ARP 2.2.

1. The first section shows that the Service is able to get an Emergency Ambulance to patients quicker than previously, and this improvement can be seen across all categories, not just the most serious cases.

2. The Service is now also able to get an Emergency Ambulance to Stroke patients more quickly, which is important given these patients require definite treatment at a hospital, and therefore the early arrival of a transporting vehicle is imperative.

3. It can also be seen that the Trust is able to meet and sustain the draft trial response standards for the Emergency Ambulance Services, both consistently and also when high demand such as winter periods and excess Heat situations, when demand spikes very quickly.

4. The number of Non-Paramedic resources have reduced significantly. Over a 2 month period this was previously 2800 (April/May 2015) verses now 471 (April/May 2017) a reduction of 83%. This ensures each patient is seen and assessed by a paramedic, providing optimal treat for each patient. The Trust now produces 96% Paramedic led Emergency Ambulance resource.

5. The number of patients conveyed to hospital has decreased as a percentage of the overall demand (accepting demand has risen). In April/May 2015 the Service conveyed 62.2% of patients to hospital, where as in April/May 2017 this had reduced to 60.57%. Using May/May 2017 activity information, this is a real reduction in patients being conveyed of 3767 in those two months. This improvement is a measure of all patients from 999calls and GP Urgent calls, whilst also counting all patients conveyed, both to A&E units, MIU's and other facilities.

	Pre ARP			Post ARP	
Clock start to a Double Crewed Ambulance on scene for all conveyed patients	Category	Avg time Minutes Seconds		Priority	Avg time Minutes Seconds
	Red 1	09:36		Category 1	09:24
	Red 2	12:42		Category 2	11:42
	Green 2	21:06		Category 3	20:12
	Green 4	39:00		Category 4	38:48
	Total	16:48		Total	15:54

Stroke Patients: Clock start to Double Crewed Ambulance on scene	Pre ARP	Avg time Minutes Seconds		Post ARP	Avg time Minutes Seconds
		13:48			12:42

Responses Per Incident (RPI)	Category			Priority	
	Red 1	1.65		Category 1	1.28
	Red 2	1.26		Category 2	1.06
	Green 2	1.18		Category 3	1.06
	Green 4	1.15		Category 4	1.05
	Total	1.23		Total	1.07

		Target		Perfromance	
Current performance	Category	Mean	90th	Mean	90th
	Category 1 R	07:00	15:00	06:49	11:24
	Category 2	18:00	40:00	10:15	18:14
	Category 3		2:00:00	18:11	38:16
	Category 4		3:00:00	33:08	79:40
	Urgent	75%		86.50%	
	Call Answering	95% under 5 seconds		96%	

Efficiency

1. Resources Per Incident (RPI), this is a count of the number resources at the scene of each incident. Pre-Trial 2015/16 – Q1 1.23 Trial (ARP 2.2) Now 2017/18 – Q1 1.07

Whilst the above can be difficult to describe, it can be quantified in the following way: The operational resource hours saved in this RPI reduction (based on 2016/17 activity)

was approximately 98,000 operational hours (typically RRV hours)
Further evidence of this reduced response rate (resource at scene) versus the activity can be seen as follows:
In 2014/15 total incident count 873,046, number resources at scene 1,107,371
In 2016/17 total incident count 942,094, number resources at scene 1,055,145
Which equates to demand growth of 69,048incidents (+7.9%) and resources at scene dropped by 49,845 (-4.5%)

2. The number of lost hours, where a Paramedic crewing an RRV has been required to travel with the Ambulance crew to hospital has been reduced by 2/3. In 2015/16 nearly 11,000hours were lost in this operational practice, where as in 2016/17 (based on Q1 data), the total losses will have reduced to 3,300hours, a further saving of 7,700hours

3. Control based staffing. Because the operational deployment of resources is much simpler, and considerably less demand is being held and managed in the Control environment, the Trust has been able to save 15 wte posts in Control (various Dispatch posts). This has been achieved through natural wastage.

4. Total Fleet mileage. In 2015/16 the Trust had an average mileage rate of 15.52miles per incident. In 2017/18 the miles per incident will be 14.7miles (average) per incident based on Q1 comparisons. Therefore, the Trust will travel 770,000 less miles per annum, which is a 5% reduction. (our latest info shows this as 15% Q1 vs Q1 data)

5. The Total Fleet asset stock has changed significantly between April 2015 and April 2017, as follows: April 2015 – 351 Emergency Ambulances and 171 RRV cars vs April 2017 – 420 Emergency Ambulances and 50 RRVs cars (being used for the small number of RRV resources, Duty Officers, HALOs and spare assets). This is real reduction in total of fleet of 52 vehicles (-10%), whilst at the same time growing the actual Emergency Ambulance fleet by 69 units (+19.5%).

6. The Service has been able to reduce the overall number of response locations, where vehicles respond from. Previously the Trust had utilised a total of 130 sites across the geography of West Midlands for responding to Emergency calls, this has now been reduced by 64 locations, without affecting operational performance.

Next Steps

The Service will be able to further embed the Operational, Quality and Efficiency benefits in the coming months, these will be focused upon:

1. Ensure 100% of all resources are Paramedic crewed (currently 96%), and ensure all patients are assessed and treated by a Paramedic, utilising only 1 resource
2. RPI will be further reduced
3. Improve (reduce) the time to Hospital for key patient groups (Stroke and Cardiac)
4. Reduce the number of patients transported to hospital
5. Further reduce the number of sites where resources respond from
6. With further improvements to dispatch methods, it is intended that both fleet mileage and control staffing could be further reduced

The examples provided by trusts demonstrate that there are operational efficiencies, and all agree that this enables some reductions in cost. It has not been possible to put value on the potential savings given the level of data available, and the substantial amount of variables, such as cost baselines, performance, fleet ratios and other factors.

The most significant demonstration of the efficiencies and benefits under ARP can be seen in the response from West Midlands Ambulance Service (WMAS). This trust has been able to demonstrate improved quality, for example response to stroke and improved performance across all categories. They have also been able to utilise the revised ARP code set to enhance their operating model, for example, by reducing the number of non-paramedic resources and as a result manage a reduced conveyance rate to the Emergency Department. The efficiency demonstrated by this trust is significant with Resource Per Incident (RPI) reduced to the point where they estimate an operational benefit of 98,000 hours annually, and the management of activity growth with a reduced resource at scene requirement. They have also been able to reduce numbers of control room staff as the number of outstanding incidents within the dispatch system has been reduced. WMAS have been the only trust able to identify fleet and fuel benefits with mileage reductions per incident and fleet stock reducing. WMAS have also been able to start work using ARP that has the potential for reductions in estate and associated costs.

It is clear from the YAS and SWASFT updates that the same benefits have not been demonstrated or identified as attainable. These trusts have come from a different starting

position, particularly in relation to performance and fleet mix. Both have stated that their starting position incorporates a higher ratio of solo responder vehicles compared to double crewed ambulances. It is evident that to operate efficiently under ARP the opposite is required. Subsequently, the financial impact of ARP will require an initial capital investment in fleet, with associated costs in depreciation and fuel. In addition, there would likely be a recurring financial impact as the workforce increases required to resource double crewed ambulances are greater than with solo resource. As activity grows this financial impact grows proportionally.

The presentation of this information is informative and with further work should enable a true financial impact to be completed. However a single impact assessment would likely only be applicable to a single trust at a given point in time due to the variation in baseline positions. A financial impact assessment therefore needs to be completed on a trust by trust basis, ideally following a nationally determined process with nationally determined measures to ensure consistency.

Conclusion

For the most acute category of patients (Red 1), response times did not improve under phase 1, and while the call categories are not directly comparable, the same appears to be true under the coding trial. However they have not deteriorated and there are likely to be several factors which influence this, such as demand and resource. The performance improvement has been in the Red 2 category and given the operational efficiency which is evidenced this can be viewed as an actual improvement with associated benefits to patients in terms of speed of response and the assumption that this is likely to improve outcomes.

Phase 2 is more difficult to assess because of the significant model changes. The results do not show improvements to the actual performance figures as they remain stable. The fact that these remained stable during a period of winter and increased demand for most trusts does show that phase 2 is likely to have had a positive effect on performance. Under the current model the trends have generally shown deterioration in performance in the winter, and this has been the case for the non-pilot trusts over this period of time.

The operational efficiencies are more defined than the financial. The results demonstrate reduced stand downs and reduced multiple allocations to incidents, with the most significant benefits shown within WMAS. This supports the case for change as ambulance trusts are better placed to improve quality of response, clinical outcomes and deal with increases in demand without the current investment requirements. This can be of particular benefit in managing delayed responses reducing long delays.

There is clear opportunity through further development of ARP, and current wider work streams within the Ambulance Improvement Programme, to deliver the recommendation of the National Audit Office to set a standard operating framework for ambulance trusts. Consideration needs to be given to the metrics and indicators which are used to measure successful delivery to ensure consistency and focus on clinical outcomes.

There is evidence in the analysis that operational efficiency has been achieved and there is an assumption that financial efficiencies follow this, but without full costing information this cannot be definitively assessed.

The variation in realisation of benefits between trusts shows that there are several factors which need to be taken into consideration. The most significant being the starting or baseline positions of the trust. For a performing trust pre-ARP which has an operating model close to that required to deliver under ARP it becomes a clear enabler to significant improvements in quality, operational efficiency and cost. However for a non-performing trust with significant change requirement to deliver the operating model, in fleet and workforce, ARP has a cost implication. It becomes an enabler to improvement but does not allow any actual cost reduction.

In terms of clinical outcome the results are essentially inconclusive due to the volumes and time period over which they have been assessed. The downward trend in the ACQIs seems to be reversed in the latest dataset released by NHSE however this is over only a three month period and further analysis is required to understand if this can be sustained and improved. Without evaluation of patients conveyed to hospital throughout their whole care pathway it is not possible to state that the change materially improves a patients care and outcome, as the majority of ACQIs are time based standards rather than outcome based.

The financial efficiencies under ARP are assumed and not properly evidenced in this impact assessment; however the narrative provided by trusts indicates that cost avoidance has, or at least has the potential, to be realised. The fact that there are operational efficiencies evidenced supports the assumption of financial efficiency and the evidence is so strong, particularly in the case of West Midlands Ambulance Service, that it should be considered as tangible. It is not clear whether the cost benefits are enough to offset some of the cost implications, but it is assumed that in the longer term this could be the case. The outcome from this is a degree of activity growth and/or quality improvement can therefore be managed through the efficiencies rather than investment.

This information does not allow full commissioner understanding of actual cost or financial impact of ARP that would allow commissioning decisions to be as informed as they are for

other NHS healthcare decisions. It is clear that local modelling has been completed, or is planned, for most trusts which begin to explore this. It is likely that each provider will need to complete this individually with their commissioners, however for consistency this will require national direction and oversight.

When considered against the 'do nothing' option it is clear that the current model is not fit for purpose, does not support efficient behaviour or use of resource and needs to be reviewed. As suggested by the ScHARR analysis the Ambulance Response Programme provides a viable enough option with strong evidence of operational efficiency, financial efficiency in terms of cost avoidance associated with this and potential improvements to clinical outcome to support the case for change. However it would be prudent to implement with the options for further development retained, as the programme is rolled out and the wider impacts more clearly understood.

The impact on local commissioning is not completely clear from this assessment. Commissioners and providers, through STPs will need to put in place a range of measures to understand the local impact of ARP and the requirements on the rest of the system to best support the outcomes.

Written by:

Richard Crocker
Associate Director – Provider Management, Urgent and Emergency Care
South Central and West Commissioning Support Unit

Reviewed and edited by:

Liam Williams
Director of Performance
South Central and West Commissioning Support Unit

References

1. National audit office. NHS England – NHS Ambulance Service. Report by the controller and Auditor General. NAO, January 2017.

https://www.nao.org.uk/report/nhs-ambulance-services/

2. House of Commons. Committee of Public Accounts – NHS Ambulance Services – sixty- second Report of Session 2016-17

https://www.publications.parliament.uk/pa/cm201617/cmselect/cmpubacc/1035/103502.htm

3. Ambulance Response Programme – evaluation of Phase 1 and Phase 2 – draft final report. J Turner et al. June 2017

Appendix 1 - ScHARR Indicators

1. Percent of Red 1 incidents with resource on scene within 8 minutes

2. Percent of Red 2 incidents with resource on scene within 8 minute

3. Percent of Red incidents where a conveying resource arrives within 19 minutes

4. Percentage of DX014 Red incidents

5. Red Incidents – Median Time to Treatment

6. Percent of Incidents resolved by Hear and Treat

7. Red 1 – Average allocations per incident – All resources

8. Red 2 – Average allocations per incident – All resources

9. Green 2 – Average allocations per in incident – All resources

10. Red 1 – Average allocations per incident – Core Resources

11. Red 2 – Average allocations per incidents – Core Resources

12. Green 2 – Average allocations per incident – Core Resources

13. Red 1 – Average responses on scene per incident – All Resources

14. Red 2 - Average responses on scene per incident – All Resources

15. Green 2 - Average responses on scene per incident – All Resources

16. Red 1 – Average responses on scene per incident – Core Resources

17. Red 2 - Average responses on scene per incident – Core Resources

18. Green 2 - Average responses on scene per incident – Core Resources

19. Red 1 – Median time from call connect to resource allocation

20. Red 2 - Median time from call connect to resource allocation

21. Green 2 - Median time from call connect to resource allocation

22. Red 1 – Median time from call connect to resource on scene

23. Red 1 –Time from call connect to resource on scene – 95th percentile

24. Red 2 - Median time from call connect to resource on scene

25. Red 2 - Time from call connect to resource on scene – 95th percentile

26. Green 2 - Median time from call connect to resource on scene

27. Green 2 - Time from call connect to resource on scene – 95th percentile

28. Red 2 – Clock Start triggers

29. Hear and Treat re-contact rate

30. See and Treat re-contact rate