Improving out of hours care: what lessons can be learned from a national benchmark of services?

Reflections and recommendations for commissioners and providers of out of hours services in England based on the first two rounds of the benchmark in 2009

January 2010

Contents

Executive summary

The Primary Care Foundation has been privileged to look in great detail at the operation of out of hours providers in over 80 Primary Care Trusts in England as part of the benchmark of out of hours services. We have now completed two rounds of the benchmark with reports to PCTs and providers in March and November 2009. In doing this we have come across a number of lessons for providers and commissioners. This document takes on the ambitious task of describing these lessons so that services can benefit from them. Many will apply to only a small number, but we would be surprised if any service will not find something of value in this report. In addition to making the document available to all participants in the benchmark we will also make it available more widely and have provided the draft version to the Department of Health to support their internal review of out of hours services.

Whilst we have included the detail that we hope is necessary for the provider service to understand how they might fix a specific issue or improve in some area of their operation, the report is aimed primarily at commissioners. This is not just because they are our clients who pay for the benchmark, but rather because it is their role to ensure that out of hours services (whether part of the PCT, independent mutual organisations, commercial providers or any combination) are part of the "self-improving NHS". This goal is not just at the heart of the competencies for world class commissioning but was also central to what Scally and Donaldson highlighted in their definition of clinical governance as the process of *"continuously improving the quality of services and safeguarding high standards of care"*.

The report identifies a number of key points that we have attempted to summarise below. Importantly though, we emphasise that because of its purpose in identifying opportunities to improve individual services, the report has to focus on problem areas. The reader should not assume that the faults described apply to many services - but we expect and hope that all services will learn of some opportunity to improve from the detail that we include.

Key observations and lessons include:

- Out of hours services are improving. Most provider organisations have had to make a rapid transition from what might, perhaps harshly, be described as 'rota organising clubs' into true healthcare providers. In doing so they have got much better at:

- Matching capacity to predictable demand, giving ample time for clinicians to do their work well
- Meeting performance standards
- Introducing governance processes to ensure a consistent and safe response to patients
- Engaging local clinicians in the service.

Although we describe how out of hours services are held to account on a much broader range of important measures than most primary care services we identify that there is an opportunity for many services to improve further.

- Patients value a responsive service and associate this with good care. We provide evidence of the close correlation between these two factors and demonstrate that the public recognise the difference between responsive services which are rated highly on the care provided and the helpfulness of the health professional. We also highlight how wide the difference is between the responsive and the comparatively slow.

- There are a number of different models for out of hours provider services. We describe these, identifying the adverse impact on patients and on reporting where appropriate. We highlight particularly the consequences in some PCTs where the service is split between two providers, where the service is integrated with other types of service and where double assessment or misallocation of case type takes place. In nearly every example where we have come across this the service is less good for patients than was being regularly reported to commissioners. We also describe how some of the processes for referral to the ambulance service or for call-streaming seem to us to fall short of the rigour and robustness of others. We make detailed recommendations that are about addressing the adverse impact on patients and the weakness in reporting associated with some of the minutiae of the process.

- Many providers are falling short on the standard for definitive clinical assessment of urgent cases which we see as an important issue of patient safety. We also highlight the enormous range across different services in the proportion of cases that are identified as urgent and suggest that commissioners should look particularly carefully at those that are well below the norm. We suspect that potentially urgent cases are missed. In this section we provide a considerable amount of detail about performance against the main standards and the variation in reporting. The examples range from the simple number of cases that are counted through to examples where the

difference between what is reported and performance against the quality requirement is very significant.

- We use some of the detailed slides from one service (shared as part of the feedback events following the first round of the benchmark) to highlight just how much variation there is within a typical service between individual clinicians. They demonstrate that the response will often be shaped more by who deals with the case rather than the details of the case itself. We describe some of the elements of the governance processes that we feel should be in place if the service is to manage this variation and work towards a consistent, safe and appropriate response. A crucial element of this is that it does not just involve looking at the outliers, but involves consistent feedback to individuals comparing them with their peers so that they can identify specific things that they might do differently for the benefit of patients and the service.

- We have looked at the cost of the different services and plotted it against population density and the volume of calls. There are striking variations, with some areas receiving three or four times as many calls in relation to their population than others and with services which have very similar demand and geography costing very different amounts.

- We have suggested three criteria to arrive at a small number of 'good all-rounders'. These criteria are based on objective measures. Although we may modify and develop the criteria we feel that identifying the 'good all-round' services and their cost will help commissioners judge value for money. Our conclusion is that those that perform well on all these factors are far from being the most expensive, but also that the very cheap providers do not appear to have the management headroom to perform consistently enough to feature in this group.

- Throughout the report we make recommendations (23 headline recommendations with considerable detail below) many of which are addressed to commissioners. However we have also included a section that specifically looks at commissioning. We urge commissioners to make use of the enormous amount of detail available about their out of hours service. We suggest that they use this report as a check-list and we urge them to be less timid in managing providers in the period between tenders. We firmly believe that it is the role of the commissioner to ensure that the service provided is sound and safe. Patients deserve it and rightly expect no less from them.

- Finally we describe some of the developments in the benchmark that are already taking place, how with the help of the user group we will develop it

further to look more widely at integrated services (as fast as
provider coding and PCT costing allows) and how we are already looking at
enhancing the benchmark to include aspects raised by users and by the
current Care Quality Commission enquiry into Take Care Now. We also outline
the plans for a detailed review of performance over the Christmas period and
the next benchmark (with patient feedback) planned for mid 2010 which we
expect to be open rather than anonymous. We are confident that this greater
openness will help both commissioners and providers to improve their
services still further. We also remind services that we are planning a detailed
session for analysts and those involved in reporting to help them to
understand any of the detailed points in this report that are not clear.

In our conclusion we make one more important point. We would be delighted to see
this report, the detailed feedback and the report on each service used to improve out
of hours services. We are happy to play our part in this process and to participate in
conference calls to discuss specific issues - contact details are included for this
purpose. In the end, however, it is the responsibility of commissioners to ensure that
the wealth of comparative performance data and analysis is turned into practical
action for improving care for patients.

Introduction and background

Purpose of the report

The aim of this report is to help commissioners and providers identify how they might improve their out of hours service for patients, providing good value for money as well as good care.

The out of hours sector has consistently richer, more accessible and comparable information than any other area of health care. The national quality requirements cover a broad range of measures, and performance of services in this sector is improving steadily.

We occupy a privileged position, with out of hours services providing us with data and detail about their operation, sometimes seeking advice in how to address specific weaknesses that they recognise. We are grateful for the assistance of the many services that have worked with us to provide information for the benchmark and our hope is that this document will be useful in helping them to develop still further. Our intention is to try to describe the lessons for providers and commissioners. The document is to be made available to participants in the benchmark and more widely. The first version has also been made available to the Department of Health in support of their internal review of out of hours services.

The Primary Care Foundation has now conducted two benchmarks of out of hours services. In the process, we have looked in detail at the performance, cost and operational processes of a considerable number of services. There are some valuable lessons and insights that we feel should be shared with commissioners and providers of these services. We have run nine feedback sessions in different venues that were intended to allow services to understand the benchmark comparisons. Although we mention some of the comparisons in this report our focus is different. The intention here is to describe many of the lessons that seem to us to be important ones for commissioners and providers. Some only apply to a small number of services, but others apply more widely.

Because we are highlighting the lessons it is necessary for us to describe some of the failings or shortfalls in some services. In a few cases the risks are obvious. The reader should be careful not to assume that the failings apply to all or even most services. On the other hand, we would be surprised if there was a single service that cannot learn something from the detail that is included within this report.

Background

Out of hours services provide primary care services to patients from 18.30 to 08.00 during weekdays and throughout the day and night at weekends and bank-holidays. The service is most commonly accessed initially by telephone which may lead to telephone advice, inviting the patient to a care centre or to a home visit.

In 2004 most GP practices passed formal responsibility for 24 hour care of their patients to their Primary Care Trust, though a small number elected to keep this responsibility. Since nearly all GPs across the country were already involved in a variety of arrangements to commission the service from others or to share the workload through a commercial, mutual or co-operative organisation, patients were seldom seen by their own GP before this change took place. However, in many areas new provider services were appointed, and across the country many GPs, no longer required to provide cover overnight and at weekends as part of their contract , chose to stop working sessions during the out of hours period.

The benchmark

The benchmark that the Primary Care Foundation carries out uses data from a variety of sources.

- A web-return completed by the commissioning PCT and signed off by the provider service that describes the scope of the service commissioned and the cost.
- A web-return completed by the provider giving detail on such aspects as the telephony performance, volume of calls and governance processes. Nearly all of this is visible to the PCT.
- A data extract from the system looking at four separate weeks spread so as to provide a representative sample of the normal performance. Typically such extracts cover several thousand cases (sometimes many more) and the use of this extract allows us to be sure that the comparisons are, as near as practical, like for like.
- A survey of patient experience conducted for us by CFEP which complements the Ipsos MORI GP survey that we have also drawn upon to provide comparative information.

The benchmark is detailed and thorough. We have been able to build on earlier work by the National Audit Office, Audit Commission and Healthcare Commission and have attempted to make sure that, for example, costs are adjusted to ensure that an

allowance is made for those services where the headline cost is not equivalent to the arms length contract cost because the provider service receives support from the commissioner who provide such services as financial reporting, HR support, IT support or provide building and cleaning services that are not included in the nominal cost. In addition the benchmark allows time for providers and commissioners to validate information before the comparisons are made.

We take great care to ensure that the comparisons are as accurate as they possibly can be. The process not only makes the data entered by provider and commissioner visible to the other it requires sign off of some information. In addition we carefully feed back the findings from the data analysis to PCTs and providers before the comparison is made. Every service is given every opportunity to ensure that the information is correct. Those that have chosen to opt out of this process have only themselves to blame if any detail is incorrect.

The first benchmark results were reported to services in March 2009 whilst the second was reported in November 2009. In both cases the aim was to look at the normal operation of the service during weeks without bank-holidays. The plan is that the next benchmark will look in more detail at the service during the very busy Christmas period when out of hours services form a crucial part of the urgent and primary care service as they are the prime point of contact for patients.

Structure of this report

Within the next section of the report we have described some of the significant progress that has been made by the out of hours sector over the last five years. Thereafter we look at various facets of the service and, within each section, we have chosen to give an overview of the topic and to include our findings and comment using headings that emphasise the key points. We end each section by describing the lessons for both commissioners and providers. We have also chosen to sign-post any proposed development of the benchmark where that is relevant.

Where any detail is included about coding or the system we have used terminology and field names from the Adastra system because it is used by the vast majority of out of hours services. We are confident that the small number of users of other systems will be able to identify the equivalent field or apply the general lesson that is relevant to their system.

Progress made

With the changes made to the GP contract in 2004 the government hit the out of hours sector with a classic 'double whammy'. It would have been hard for the sector just to handle the change because so many doctors (who had previously felt obliged to participate in the rota because of their 24 hour responsibility for patients) decided that they would stop participating in their out of hours rota. This difficulty was compounded because contracts were put out to tender and there was a consequential change in management and a loss of continuity. Although TUPE applied, much of the knowledge in the co-ops and other organisations that were previously providing the service was held by the GPs that ran the different services, and in many places the key individuals were not in place when the new contractual arrangements went live.

We looked at a considerable number of services in 2005 and 2006 and it was not uncommon to find the service massively over-stretched at busy times, with performance well below the national standards. Call-handlers were often involved in comfort calling patients because of the extended delays and, In some cases, a vicious spiral was in danger of developing where clinicians were unwilling to work in an environment that they felt was unsafe because they had no time to properly assess and treat patients - but as they stopped working at those times the situation was made worse.

The reader should not assume that all was well before 2004. Although many of the services were adequately staffed because of the obligation in many co-ops for each doctor or practice to participate there were many short-comings. We have no reason to believe that the standards of timeliness were consistently met in any service, arrangements for governance were weak, little use was made of the information available and the quality very much depended on the individual clinicians working at the time.

PCTs and providers have got through these challenging times. Some managed to retain sufficient clinical and other staff so that they suffered only a little, but others have had to learn, not just how to run the operation as previously, but how to drive up performance whilst encouraging clinical staff to participate.

The sector is competitive. Most PCTs have put the service out to tender at least once, and some more than once, since the contracts were first let at the end of 2004. However it is competitive not just because of the tendering process, but also because it is a sector which is rich in information and where there are clear quality

requirements. Increasingly, there is also good comparative information about what patients think about the service.

The combination of these factors means that provider organisations have had to grow up. It would perhaps be only a small exaggeration to say that they have developed from being a co-operative club (where the main focus was on sharing the workload evenly and the costs appropriately) into a proper healthcare provider, with sound governance arrangements, good management and measurement. In many places there are now more than adequate numbers of clinical staff available to fill the rotas because the local clinicians have confidence in the service and wish to be part of it. The strength of the sector is perhaps demonstrated by the key role that many services played in dealing with the surge in demand for primary care advice because of the summer swine-flu. Not only did they provide much of the response during the out of hours period, but many also played a valuable part in setting up local arrangements for call-handling and the distribution of Tamiflu.

Despite this progress there are opportunities to do things still better. This document focuses on these opportunities and highlights the lessons that we hope will allow services to improve care for patients, supporting improved effectiveness and efficiency still further.

Patients' views

Overview

We think it is important to include our analysis of the patients' views of out of hours services. Many services have a high proportion of respondents rating the care provided as good or very good - but the range from the best to the worst is significant (78% to 46% at PCT level) with an average of 66%. We would have hoped to find a greater proportion above 65%.

It is no surprise that patients rate care from out of hours services lower than they do that from their GP practice where the patient and clinician are more likely to have a long-standing relationship. The survey asked a series of questions about different aspects of care provided by the patient's practice and the typical proportion rating the care as good and very good about both nurses and doctors was 86%.   level..

There are two good sources of information that we have for the second benchmark which provide a valuable way to look at the treatment and care received by patients. As part of the benchmark we commissioned a survey from CFEP for which over 9000 replies were received which provides detail that complements the Ipsos MORI GP survey where, among the total of 2.1 Million responses, close to 300,000 answered the questions about out of hours services. The Ipsos MORI data is available by PCT (and by provider within PCT) through these links http://www.dh.gov.uk/en/Publicationsandstatistics/PublishedSurvey/GPpatientsurvey2007/index.htm and http://www.gp-patient.co.uk/200809results/ . Importantly the survey is running to a cycle so that the information will be refreshed for two quarters in each year.

Overall the majority of patients are pleased with accessibility by phone, with 79% saying that it was easy to contact the service.

Response times matter to patients

It is perhaps unsurprising that there is a relationship between the patients' perceptions of the speed of response and their view about the quality of care received. The vast majority of cases that present to out of hours services are minor self-limiting conditions and require relatively quick consultations with a clinician,sometimes over the phone and sometimes face to face. In these circumstances the patient will perhaps consider a speedy response to be of greater importance relative to other factors compared with if they had a number of

interactions, for example with a hospital or a GP over an extended period.
Even with this expectation, it is still a surprise that the correlation in figure 1 is so close.

Figure 1 – comparison of patient perceptions of how quickly care was received against their rating of the care received (Ipsos MORI GP survey)
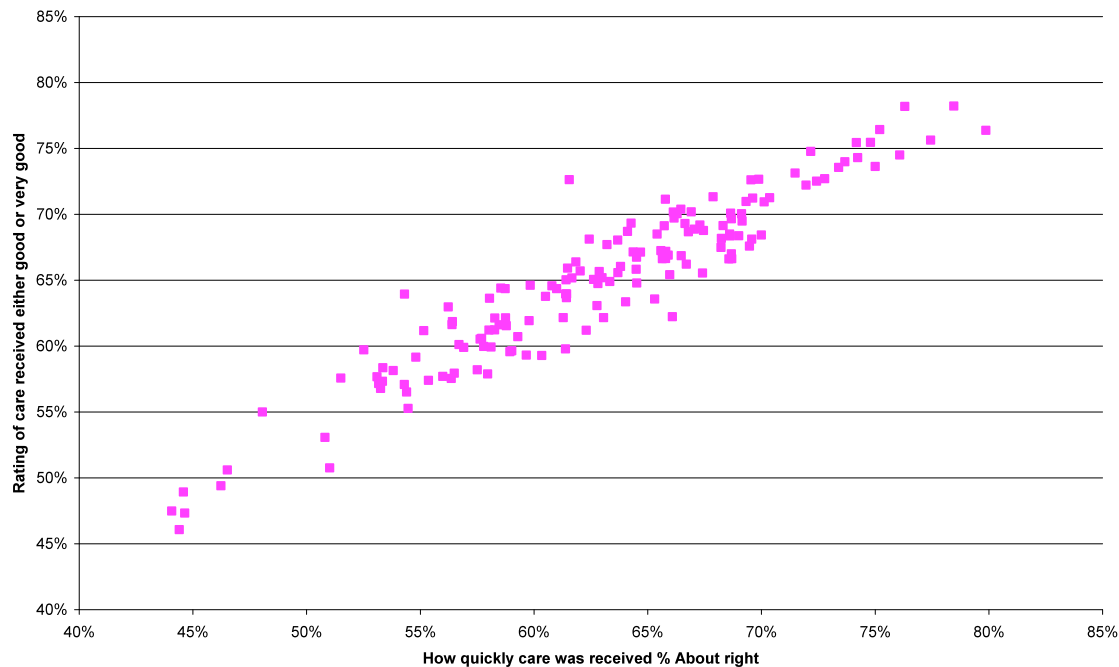
Each point in the graph represents the views of close to 2000 patients from the PCT about the out of hours service from which they received care. There is a strikingly close correlation between the overall evaluation of care and the patient's assessment of timeliness. However there are also large differences between services with only 45% of patients in some areas rating the care as good or very good whilst in other PCT areas the figure is over 75%. The lesson for providers and PCTs is that if patients perceive the service as responding quickly enough, then they also think it provides good care.

Using the benchmark data for time to definitive assessment we have looked at the relationship between the responsiveness of the services and patients perception of how quickly the service responds and their view of the timeliness of response and the care received. There is a clear relationship between lack of responsiveness and the patients' view that it was too slow.. This carries across to the perception of the patient about the care received as measured both in the Ipsos MORI data and in the more detailed questions from the smaller but more detailed CFEP survey that was commissioned as part of the benchmark.

Figure 2 showing the inverse relationship between the time to definitive clinical assessment and the numbers that felt the service was too slow.
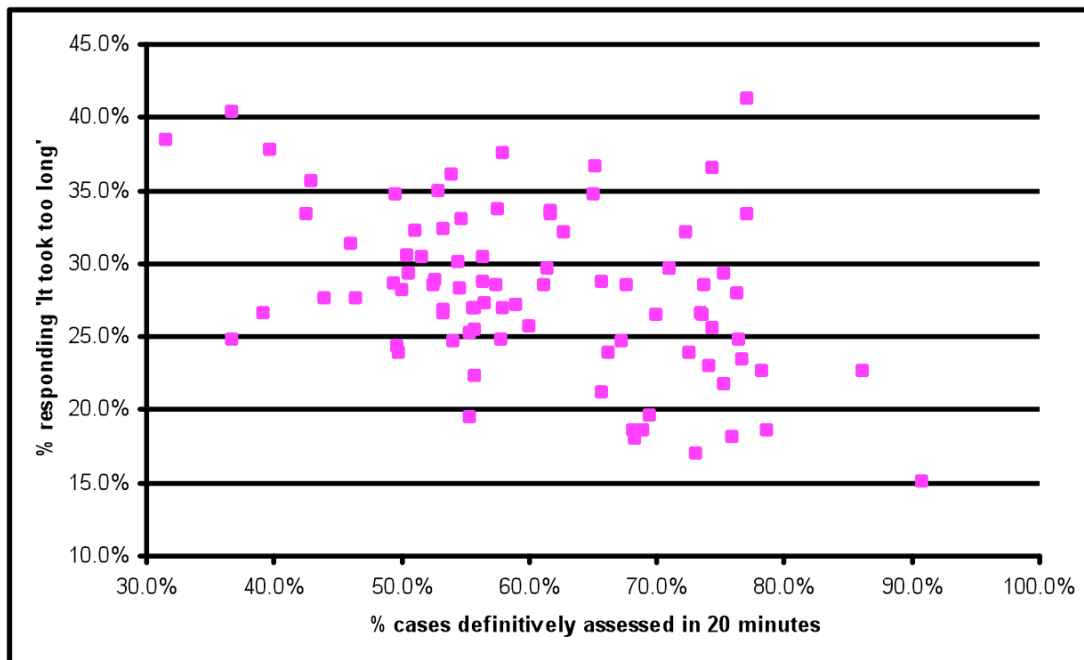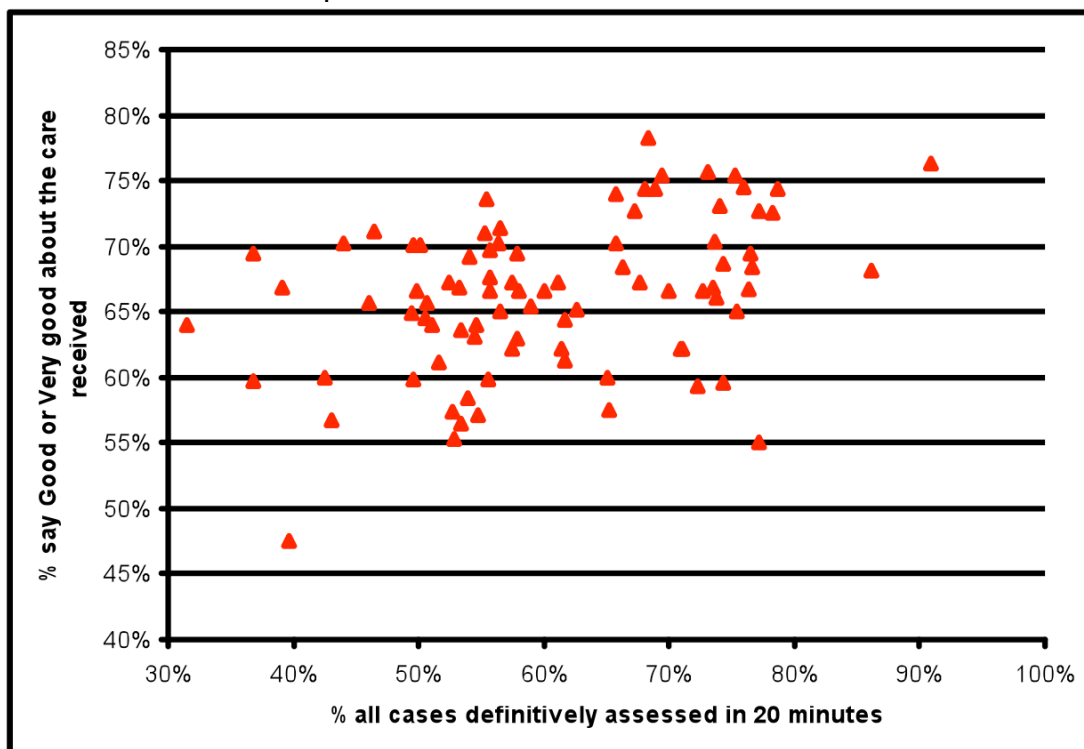

Figure 3 showing the relationship between the time to definitive assessment from the benchmark and the patient view of the care received

Lessons for commissioners and providers

There is an obvious lesson for both commissioners and providers.

Responsiveness matters to patients

Speed of response appears to be far more important to patients than the other differences that are found between services. One example illustrates this point:

Many PCTs and providers spend a lot of time looking at the split of dispositions (telephone advice, attending a centre receiving a home visit). The argument has often been proposed that services that see more patients face to face are offering a better service that is valued more highly by patients.

We looked at two services that were very different in this respect. One advised 73% of patients by telephone and saw only 2% at home whilst the other advised only 37% by phone and over 15% received home visits.

Despite their very different approach and styles of operation both of these services were very highly thought of by patients. They were both in the top groups as measured by the Ipsos MORI and CFEP surveys in respect of their responses to the questions about the care received, the help received from the health professionals and the appropriateness of the response (whether the patient felt they received the right disposition).

There was one common factor – both services are extremely quick in responding to patients by phone and in seeing them face to face.

Our conclusion is that those services that do not complete definitive assessment of well over 60% of all cases within twenty minutes and that do not see patients face to face promptly will never score highly with patients.

Within the benchmark we will continue to monitor services by looking at the percentage of all cases definitively assessed in 20 minutes and will, in future benchmarks, look at those cases that end with a face to face consultation and measure the proportion that are completed within two hours from the initial contact. We see that these measures will augment the National Quality Requirements and focus on an aspect that clearly matters to patients.

Commissioners should use the Ipsos MORI data

Although we did not specifically ask every service about it, we found none that had used the Ipsos MORI survey data to look at how patients see their service – indeed most commissioners and providers seemed to be unaware that the data was available.

The survey goes to a very large number of patients. In the first round a larger than normal number of responses were received so that around 2000 patients per PCT had answered questions about the out of hours service that they had used. Now that the survey is into its quarterly cycles approximately another 1000 responses are being added to the data in two quarters each year. The survey also asks about a number of other aspects, such as how easy it was to access the service by telephone and about the ease with which medicines could be obtained.

With such striking differences between services, this information could be used to evaluate the service, to monitor it and to set targets for improvement. If it is the case, as it appears, that this information is not being used to assess the performance of the provider and as part of the information about prospective bidders when services are put out to tender it would seem to be a neglect of the commissioning responsibilities of the PCT.

With the inability of many services to report in line with the telephone standards (see below) we may use the patients' perceptions of the ease with which services can be accessed as the best comparative measure of the telephony response for future benchmarks.

Variations in operating model

There are significant variations in the operational model of service used in different services. We think it important to highlight these, describing some of the advantages and issues associated with each before we get into some of the more detailed discussion – especially because the models can have a major impact both on what is measured and reported and on the service to the patient. We have described the conventional model first and then looked at the variants on that model and the different operational models together with the features and the advantages or risks that need to be managed in comparison with the conventional model.

The conventional model

The standard model of service is based on expecting nearly all patients to telephone the out of hours service, although a small number will also walk in without a preliminary telephone call. The process is that the call is received by a call-handler who will take the essential details of a telephone number, name, address, details of the patient's own doctor and a summary of the condition that the patient is calling about. The system normally provides support in this process by identifying previous patients and addresses of those calling from that phone number or using the post-code to speed up capture of the address details.

Importantly, the non-clinical call-handler is expected to identify if the condition is potentially life-threatening and to assign a priority to the case. To do this they will either use a paper-based protocol or a clinical decision support system within the software. In either case it is important that the training is such that the call-handler is alert for the potentially urgent or life-threatening condition (that may not always be well described by the caller), but is also not over-prioritising cases so that excessive numbers of patients are sent to hospital or identified as urgent. In most cases, if the call is immediately life-threatening, the call-handler will ensure an ambulance is called, but in some cases some services will check with a clinician before this is done.

The call is normally completed by the call-handler and passed across into a queue for a clinician to ring the patient back.

The clinician will pick patients from the queue, ensuring that urgent cases are rung back earlier but generally picking the oldest case first. When they ring back (and after checking that they are speaking to the patient or carer that rang), they will assess the case and either advise the patient over the phone or arrange a face to face appointment. Telephone advice is usually for self-care, with suitable safety-netting about what to do if the condition gets worse or does not respond, but will

sometimes involve referral to another part of the NHS such as the hospital, the district nursing service etc. The face to face consultation will either be in the care centre, in which case the clinician will make the appointment for the patient, or a home visit will be arranged, in which case the dispatcher or driver will usually telephone with an estimated arrival time.

Minor variations on the conventional model

There are a number of variations on the conventional model that are worth commenting on because they affect the process, the information reported or care for the patient.

Calling an ambulance

When a call-handler identifies that a patient requires an ambulance there should clearly be no delay before the ambulance service is notified. However there are a number of different processes that we have come across including:

1. Checking that the patient is able to do so and advising them to ring 999 immediately
2. Transferring the call immediately to the ambulance service
3. Keeping the caller on the line whilst they advise the ambulance service of the various details confirming to the patient when the ambulance is dispatched
4. Involving the ambulance service in a three way call so that the necessary details are collected by both services and handing the patient over to the ambulance service.

Three things seem to be important about how this process operates:

➢ That there should be minimal delay in dispatching the ambulance
➢ That sufficient information should be gathered to meet the needs of both services – which in the case of the out of hours service includes identifying the patient and their GP so that the practice is notified the next morning
➢ That the two services should already have agreed the process and protocols to minimise delay when an ambulance needs to be despatched.

Processes 1 and 2 from the list above fail these tests. Process 1 always implies a delay whilst the patient repeats many of the details already taken by the out of hours service and will certainly involve a repeated assessment (with the risk of the ambulance service reaching a different conclusion and referring the patient back to the out of hours service). Process 2 may result in the out of hours service not getting

all of the details required to identify the patient and their GP. For these reasons we recommend PCTs promote dialogue between the out of hours provider and the ambulance trust and that they agree a process that allows the out of hours service to collect all of the necessary details to identify the patient and the information necessary for the GP without delaying despatch of the ambulance.

In our recommendations about this process below we have also suggested that services should look to ensure that the ambulance service and (probably through them) the hospital, receive details of the patient's history (including relevant earlier consultations). The major out of hours systems can do this in a number of ways, though sometimes the age of the hospital system means that it is not possible to transfer this information in a way that avoids re-keying.

Warm transfer of calls

Some services have established processes whereby some calls are transferred directly from the call-handler to the clinician without the delay associated with a queue before the patient is rung-back. Where this process is established, it is most commonly for urgent cases.

There are great advantages if this process can be utilized for border-line cases that the call-handler feel might be life-threatening in that the clinician is better trained to make the fine judgment over whether or not the patient should go to the hospital and/or an ambulance is required. As described below we believe that the typical level of patients that go towards a hospital after contacting the out of hours service is between 12 and 17%. Whilst in many cases this is absolutely the right thing for the patient, this process is one of the few areas where the out of hours service can potentially reduce the referrals to hospital. The protocols used by the call-handler should always recognise situations (such as severe chest pain) where there should be no delay in getting an ambulance dispatched. However for border-line cases there are advantages in the warm transfer to a clinician. With minimal delay the clinician can then trigger the dispatch of the ambulance or suggest an alternative course of action.

There are advantages too, once such a process is established, in the warm transfer of calls in other cases. All services using a ring-back model find that sometimes, when the clinician tries to ring back, the patient has already gone to hospital. By using the warm-transfer approach, particularly for any patient that seems worried, it may well be possible to reduce the number of such patients that attend the emergency department.

Because of the obvious advantages to a patient in being able to speak to a clinician as part of their first call to the service, we have in the past looked at the likely additional cost of warm transferring the majority (perhaps 80% within one minute) of calls. We would be delighted to work with a service on testing this in practice as we our calculations indicate that the cost is far lower than many anticipate.

Nurse and doctor assessment

Many services involve nurses in the process and a number use other healthcare professionals, sometimes emergency care practitioners and occasionally pharmacists. Where nurses and ECPs are utilised in telephone assessment there are inevitably some occasions where they will find that it is necessary for a doctor to assess the patient. This sometimes occurs in those cases when the patient is unwilling to accept the advice from the nurse or ECP and the case is passed on to a doctor in the belief that he or she will be able to convince the patient to follow the recommended course of action. It is also common in the more complex cases, perhaps involving multiple pathologies, and where the patient may already be taking a number of medications when the nurse or ECP, recognizing the greater expertise and experience of the doctor will wish to pass the case on.

Whilst this will inevitably happen on occasions when nurses are involved in assessment, the variation is highlighted for two reasons

> ➢ Because of the implications of this double assessment both for the patient (who receives two ring-backs) and in reporting of the various standards for timeliness that are discussed below.
> ➢ Because, depending on the detail of the process, the number of cases that receive this double assessment can be very different. Those who have established arrangements where nurses assess all cases, or that rely on the call-handler deciding which case to send to a nurse and which to a doctor queue, seem to have more than those that put them into one queue and encourage both doctors and nurses to cherry-pick cases, with the doctors picking the more complex and the nurses choosing those where they feel that they are likely to be able to complete the clinical assessment.

Call-streaming

Some services allow call-handlers to invite patients directly to the base, usually booking them an appointment without there being telephone assessment by a clinician. The advantage of this is obvious in those cases where it is plain that the clinical assessment is unlikely to add any value or make any difference to the disposition. Two obvious examples spring to mind – one is the baby with some kind of tummy pain where it is apparent that the clinician will need to see the child, and the other is the palliative care patient where the special notes may well make it obvious that a home visit is appropriate.

However services that use this approach generally adopt it in a much wider range of cases and will often, after making sure that the condition is not urgent or immediately life-threatening, offer the patient the choice of a telephone call with the clinician or an appointment at the care centre. Patients, understandably, like this and we support its use – but only with the application of certain conditions.

Services that wish to adopt this approach should make absolutely certain that their protocols for the identification of life-threatening and urgent conditions are robust. Whilst the standard for definitive clinical assessment requires the provider to have a "clinically safe and effective system for prioritizing patients" to distinguish between urgent and less urgent cases, this is in the context of a process where even less urgent cases are assessed by a clinician within a maximum of an hour. If the process is such that the patient does not speak to a clinician until they see them face to face (potentially much later than this, allowing for time for the patient to get to the centre) then the prioritisation and identification of life-threatening conditions needs to be more robust, because the safety net of the clinical assessment that normally will take place within 30 minutes is not in place.

Providers who adopt this approach should not only have a very robust process for the call-handler prioritization, but should also make sure that this is rigorously audited, that the PCT has looked in detail at the process and has formally agreed (probably with board approval) that call-streaming may be adopted. Even with these safeguards a provider and PCT should recognise that were a significant untoward incident to arise they might be criticized if the patient had not received definitive clinical assessment within the timescales set out in the national quality requirements 9 and 10.

Finally we are sceptical of those services that only adopt call-streaming when they are busy. If the process is robust and defensable as described above, and if patients prefer it, then we would expect the service to adopt the approach at all times. Our

strong suspicion is that services that only call-stream at busy times do so
to reduce the workload and will not have trained call-handlers and clinicians in the
process well enough for the conditions above to be met.

Split services

In some places two providers jointly deliver the service to patients. The exact point
at which responsibility is split may vary. In some cases the 'front end' provider is
responsible simply for the telephone answering, only employing call-handlers; in
others they may undertake some or all of the clinical assessment; and in still other
cases the 'back end' provider is only responsible for the home visits.

Whilst the arrangement may appear to offer advantages (because one organisation
may have the skills and expertise to manage the telephony element and another the
clinical expertise) the examples that we have seen as part of the benchmark have
had a variety of problems associated with them.

Many of the issues are similar to those that arise with the double assessment when
nurses are involved – but the disadvantages become heightened by the separation of
the two organisations. We have come across cases where the process is such that:

> ➢ All, or virtually all, cases that are sent across as base consultations are re-
>   triaged by the second organisation. The patient thus has to wait for two
>   different clinicians to ring them back.
> ➢ In the case of home visits, we have identified cases where the clinician rings
>   the patient back, but the call is recorded on the system as a home visit, so that
>   there appear to have been one telephone assessment and two home visits,
>   when actually there were two telephone assessments and one home visit. By
>   extension the assumption is also that there were some cases reported as
>   home visits that were actually completed with telephone assessment This
>   abuse of the system throws out any hope of the service being able to report
>   accurately on time to definitive assessment or time to the face to face
>   consultation. Sometimes it happens so frequently that it seems to have
>   become the standard process for handling home visits.

Theoretically it should be possible to split the service in this way. However we have
yet to find a single split service that is accurately identifying when the definitive
assessment is carried out and using this in their reporting against the national quality
standard 9 and 12. The services most frequently report to the first clinical assessment
only and the delay between the two organisations is not visible in the reports from
either provider. Frequently we have surprised commissioners and providers by

reporting that the service definitively assesses less than 80% of urgent
cases in 20 minutes when they had believed it to be above 95% because of counting
to the first assessment.

Not only do split services tend to introduce delay for the patient and mean that
reports purporting to measure against the national standards are inaccurate, there is
a further problem. In every case that we have come across we have been unable to
convince ourselves that the governance and review processes are looking at the
whole service as experienced by the patient. This is often because the two contracts
have been let separately by the commissioning PCT and neither service sees it has a
duty to oversee the whole, but even when this responsibility is clear, because one
provider is a sub-contractor to the other, the clinical governance and operational
leads have often not seen it as part of the role to take responsibility for the whole
process. Without this sort of overview we do not see how the issues that we
highlighted about the overall process will ever be addressed, let alone the more
subtle, but critical, details about the decisions made by individuals involved in the
process and how these interact.

Services that are integrated with walk-in services

We applaud the attempt by many PCTs to rationalise the delivery of services by
integrating out of hours services with others such as walk-in services, minor injury
units etc. However in some places some key aspects have sometimes been lost
through integration.

The regulatory framework within which out of hours services, walk in services or
schemes for seeing patients that have attended the emergency department are
commissioned includes the need to measure and report against the appropriate
quality requirements. Patients in any PCT can expect to receive a service that is in
line with these standards. There are a number of problems that commonly occur.

The process is inconsistent with the requirements in the standards
In the first the process is such that a patient that has already been assessed by phone
and invited to the centre is treated just like any walk-in patient when they arrive.
They are re-assessed and then placed at the back of the queue (unless their clinical
need is such that they command a greater priority). Effectively this means that the
patient is disadvantaged and might as well have come straight to the centre - the
telephone call has been of no value to them, indeed it has simply delayed them
before they join the queue. Worse than this, some of the services that we have seen
judge themselves mainly against the four hour waiting time target and have not

matched capacity to keep up with predictable demand, let alone unpredictable peaks so that rather than being seen reasonably promptly on arrival, the patient is required to wait for hours in an uncomfortable environment where they are exposed to infection. This is not good care and is not an experience that is welcomed by patients.

This process not only prolongs the time for the patient and makes the service less convenient for them it also goes against the assumption that underpins the out of hours standards. This assumption is that a patient who rings and has to be seen at the centre can be advised when to arrive so that they are seen promptly (usually they are given an appointment). The process adopted in many of these integrated services not only disadvantages patients, but also (as patients will soon learn not to ring) discourages patients from ringing even when they might benefit from advice over the telephone.

The ability to report against quality standards is lost

The second issue is that in adapting two different types of services the ability to report against the quality requirements has been undermined. There are a number of reasons that this can happen such as:

- Coding is inadequate to allow the types of service accessed by the patient to be distinguished. As a minimum, we would expect the provider to be able to reliably identify and report separately on those patients: that telephoned, those that walked in, and those referred from another NHS provider (such as the emergency department). Such distinctions are important if the service is to be compared against the relevant standard. We would strongly recommend that the code used to separate these is a field set on receipt (such as the call origin or service field) not the case type. Below, and in the appendix, we provide more detail about coding.
- Coding is inadequate to identify whether a face to face assessment is an initial triage or a full consultation. Without this information it is impossible to report accurately the time to definitive assessment (QR10) or the time from the end of that assessment to the full consultation (QR12).

As a minimum, services should be reporting about each type of service separately against the relevant standards . Whilst there is no compulsion so to do, we think that there are good reasons that all cases should be reported against all the standards. We see nothing wrong with monitoring out of hours cases against the typical walk-in or emergency department four hour standard for completion and with monitoring walk-in cases against a 20 and 60 minute standard for clinical assessment and a 1 hour and two hour standard for the face to face consultation depending on the priority.

Consistent failure to meet these standards ought to raise questions about the service, even though the measure is not normally formally part of the quality requirement for a stand-alone service.

The cost of the different elements are obscured

The third problem is that often, when services are integrated, neither the provider nor the PCT has any idea about the relative cost of the different services. It is not always easy to calculate this cost, but this is no excuse for not attempting to understand the difference. To allocate the costs using reasonable activity based costing principles requires good recording, not just of the number of cases associated with each service, but also measures or understanding of which skill group is involved (which may vary depending on the time of day). Without this information it becomes impossible to compare the cost per head or cost per case of the services and therefore to know if the commissioner is getting good value for money. As always, good information supports good commissioning and proper contract monitoring.

Lessons for commissioners and providers

Calling an ambulance

Commissioners should support providers (where arrangements are not already in place) to ensure that there is an agreement in place that allows the out of hours service to transfer responsibility for a patient to the ambulance service in a way that minimizes delay, avoids any need for reassessment and that ensures that adequate details are captured to identify the patient, their address and GP so that the practice is informed the next morning. Providers should look at the detail of their process and work with the ambulance trust to ensure that the criteria described above are met. In addition they should work towards being able to transfer any relevant patient notes and details of previous consultations to the ambulance service and the relevant hospital to which the patient is taken.

Warm transfer of calls

Providers may wish to experiment with the warm-transfer of certain calls (not those where seconds may count in improving the prognosis for the patient) to clinicians with a view to providing an immediate clinical assessment. In this way it may be

possible to reduce the number of 999 calls and also reduce the number of callers who are sufficiently worried not to wait for the call back but go directly to the emergency department With this in place they may wish to extend the approach to a wider range of calls.

Double assessment

Providers that use nurses, or are involved in a split service, should ensure that they are reporting reliably the time to definitive clinical assessment (the last assessment if there is more than one), and that the reporting is not being undermined by clinicians who will telephone a patient and enter the details as if it were a home visit or a face to face consultation in the care centre.

Commissioners should ask for information so that they can monitor the number of patients and the number of telephone calls involved. Where the service is split there are often cases where there is a double assessment (one incoming and two outgoing calls), but in other cases (particularly where the systems used by the front end provider do not allow appointments to be booked), where there may only be one assessment, there can still be a third call because a call-handler/receptionist rings back to agree an appointment time. In some cases four calls (one incoming and three out-going) are not at all uncommon.

Split services

Where delivery of different parts of the same service are split between two providers commissioners should ensure that not only is there clarity over where responsibility for governance of the operational and clinical processes lies, but also that the whole process is examined to look at the decisions made by staff from both providers.

Call streaming

Where call-streaming takes place there is likely to be a longer delay before a patient speaks to a clinician, so commissioning PCTs have an even greater responsibility than is normally required for QR9 and 10 to ensure that the process is clinically safe and effective (which implies a rigorous review not just of the protocols but of the staff training) and is appropriately monitored and audited (which implies investigation of any cases which were not recognised as urgent or life-threatening on receipt and which were subsequently escalated). Approval to allow the provider to adopt call-streaming should be signed off formally at a high level.

Where call-streaming takes place, providers should still monitor and report the time to definitive assessment of those patients that receive telephone assessment or advice.

Integrated services

Where a number of different services are integrated the provider should ensure that:

- Coding is adequate to reliably distinguish the type of service - those who have phoned the out of hours number, those passed to the out of hours service by NHS D, those who walked in, those referred from the emergency department etc.
- Coding reliably distinguishes the nature of the consultation, distinguishing between an assessment and a full consultation, and identifying whether it was carried out by phone, in the centre or at the patient's home
- They report not only on the standards for that service but on the more demanding standards for all patients so that they can identify if, for example, significant numbers of out of hours patients do not have their case completed in four hours or some walk in cases that are urgent are not assessed within 20 minutes.
- Costs are attributed reasonably to each service so that the costs per case, cost per head etc. can be reasonably compared with others that may integrate with a different mix of services, and so that the value for money can be assessed.

In future benchmarks we will be comparing the performance and cost of integrated services across a wider range of measures, but in many cases we expect to find, at least until things improve, that the comparison is compromised by some or all of the issues listed.

Performance against the main numeric standards

We carry out a comprehensive analysis against standards as part of the benchmark. Within this section we have provided an overview the overall range of performance where we have been able to make comparison between services. We describe where services generally meet or exceed the standards and where they often fall short. The National Quality Requirement numbers are used as headings in both of these sections.

Comparative performance

### QR7 - matching capacity to predictable demand

Providers are expected to match capacity to predictable demand. We assess this in the benchmark by looking at the consistency with which the service responds as measured by the main indicators of responsiveness - time to definitive assessment and time to the face to face consultation (though we do not include the measure for emergency cases as it is so sensitive to one or two cases). We also look at the graph that was provided as part of the feedback to each service that showed the average time to definitive assessment for each hour of a typical weekday and a typical weekend day based on the four week sample.

Almost exactly half of the services averaged better than 90% on this composite measure of overall responsiveness. This is not a demanding standard. It is perfectly possible to achieve this level whilst failing to meet the standard for partial compliance on a number of the quality criteria. Only 15 services achieved an average that was better than 95%. To illustrate how performance drops during the busy times, we have included below the graph for one of the services that just failed to meet the 90% level and one that just achieved a 95% level. The colours indicate the proportion of cases that were definitively assessed within different time periods. As we are looking at cases of all priority, any above the green (40 to 60 minutes) by definition fails the standard for time to definitive assessment. The increased delay at the busier times in the first example indicates that capacity is not matched to demand, whilst the second example, with very few cases above the green, indicates that capacity is matched to demand.

Figures 4 and 5. Showing the time to definitive clinical assessment for two services by hour for an average Saturday or Sunday– the first falls well short of the standard whilst the second performs well, even at the busy weekend morning.
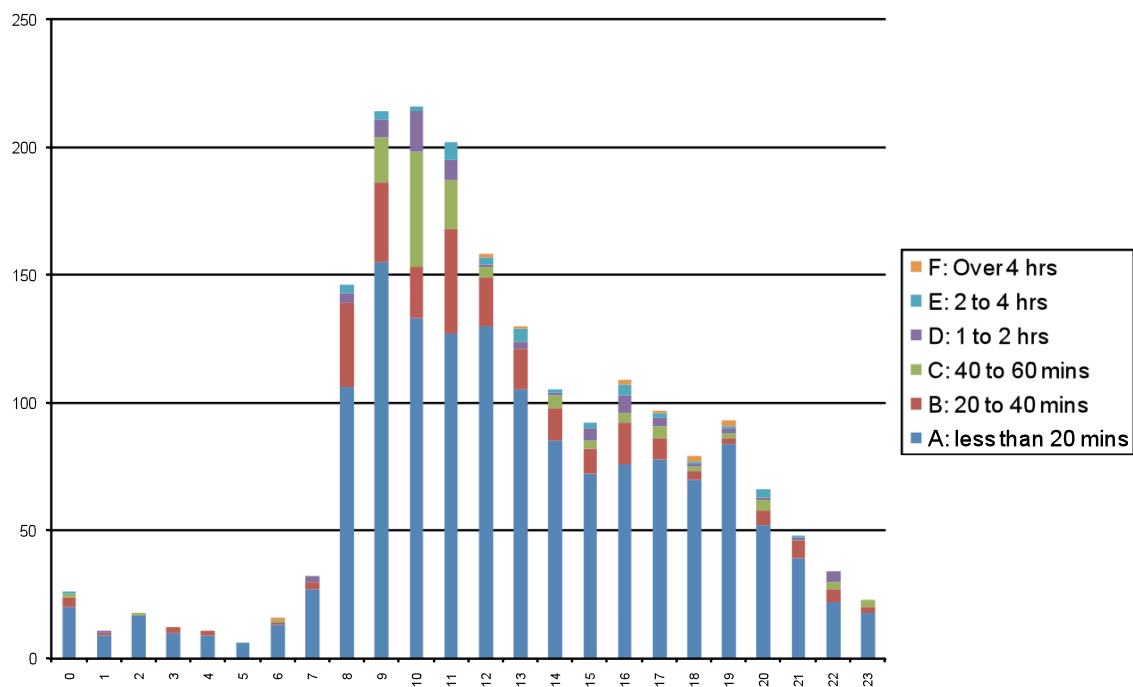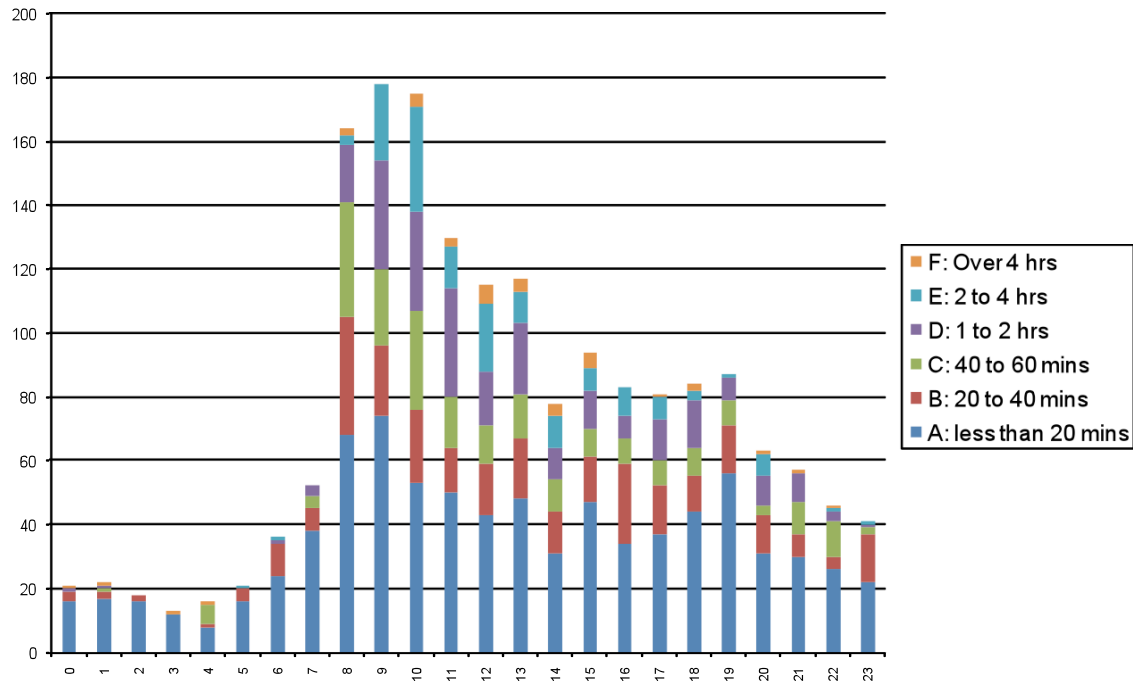
QR9 and 10 Time to definitive clinical assessment

We have described below some of the inconsistency in the measurement made by many services of the time to definitive assessment. However, the analysis completed as part of the benchmark is based on a data extract and is in line with the standard. It should be noted that we do not take account of patient attributable delay (which typically makes a difference of two or three percent) because not all systems enable this to be done and we are keen to make the comparison like for like.

The quality requirement allows services to measure the percentage of cases definitively assessed against two time periods - 20 minutes for urgent cases and 60 minutes for less urgent. Nearly all services claim to be able to *'demonstrate that they have a clinically safe and effective system for prioritising patients'* to measure their performance in this way.

However there are enormous differences in the percentage that services typically identify as urgent on receipt from well under 5% to well over 40% - and in most cases there will be much larger differences between individual call-handlers that are hidden by the averaging when looking at the service as a whole.

Figure 6 showing the enormous variation between different services in the proportion of cases identified as urgent on receipt
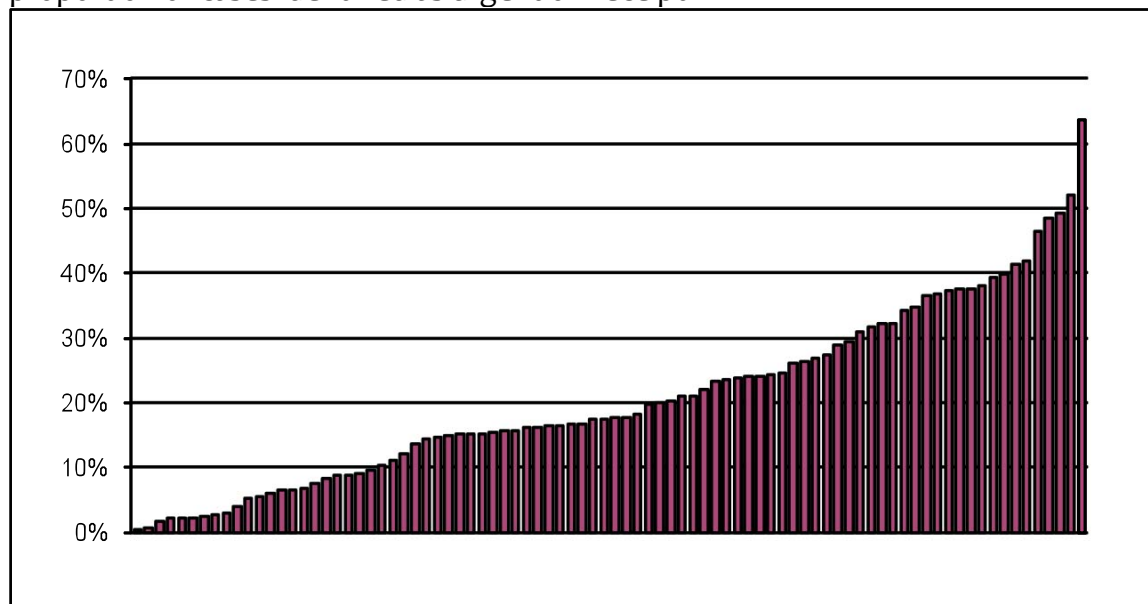
In our view, commissioners should be sceptical whether any service that identifies too low a level of cases as urgent on receipt has a process that is demonstrably *"clinically safe and effective"*. A level of less than 5%, for example, seems to us to be

so low compared to the norm that it is difficult to see how the provider can be confident that no urgent cases are missed. Unless the commissioner is satisfied that the non-clinical call handlers are reliably identifying life-threatening cases and applying an appropriate priority to all cases, the provider should be required to report against the harsher target of assessing all cases in 20 minutes.

Providers with thorough comprehensive systems for clinical and operational audit look routinely at cases where priority is escalated. Some indication that services are missing some urgent cases at receipt is given by the graph below that compares the percentage of urgent cases on receipt with the percentage of urgent cases after assessment among those to be seen face to face. Those below and to the right of the diagonal line are likely to have significant numbers of cases where priority is escalated. In the majority of cases where the percentage of urgent on receipt is less than 10%, there is escalation of priority which seems to confirm that the urgency was missed when the case was received.

Figure 7 where services below and to the right of the diagonal line show signs of significant escalation of priority compared with priority on receipt
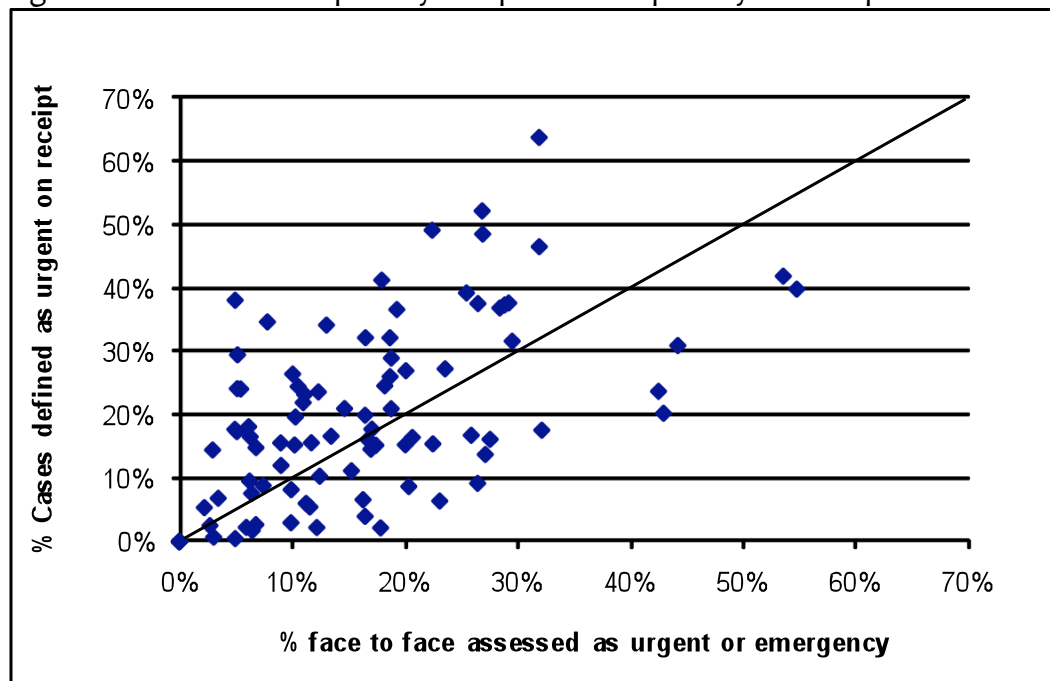
Nor should services with very high level of urgent cases on receipt be complacent. It is clearly more difficult to achieve the standard of assessing urgent cases within 20 minutes when there is a much larger proportion of urgent cases - but of greater concern, when the service does get behind (and it happens to all services at some

time because of an unpredicted peak in demand), the clinicians may have difficulty identifying the really urgent among so many.

We found few services that were fully compliant with the standard for beginning definitive assessment of urgent cases within 20 minutes of the call being answered and only 16 that achieved better than 90%, the level for partial compliance.

Figure 8 showing the range of performance in time to definitive clinical assessment of urgent cases with very services meeting the standard
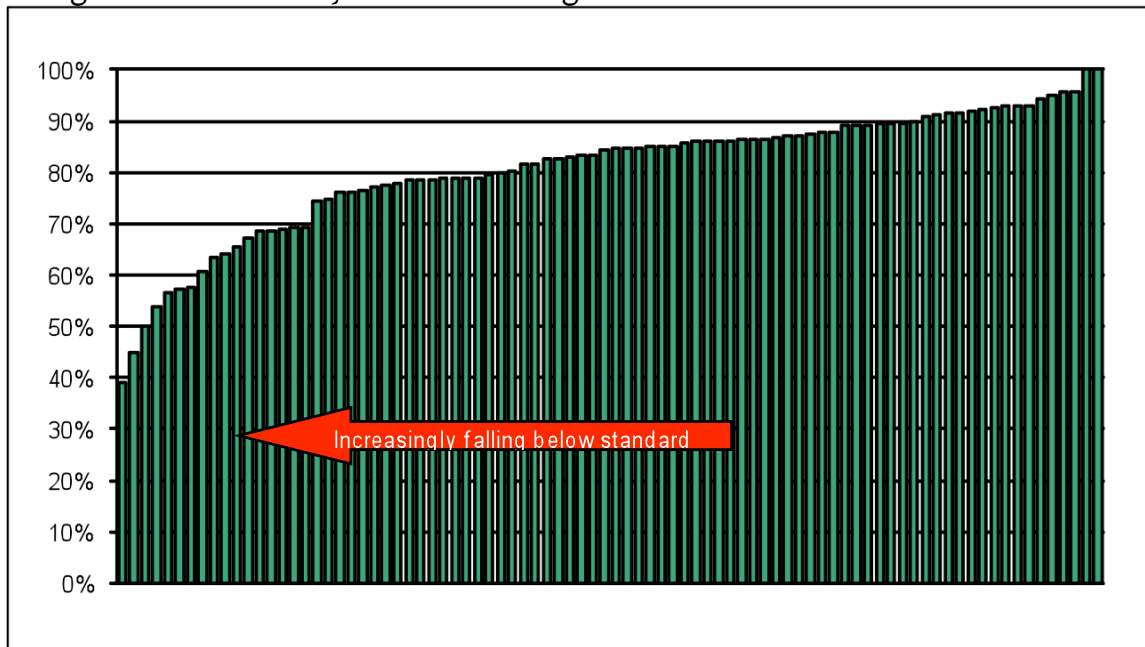
These results are not adversely affected by call-streaming. Where providers stream calls, we have excluded those cases that were streamed directly to a face to face consultation and measured the other calls on the time to the definitive telephone advice.

There are four main reasons why services fall short on this measure.

- Staffing is inadequate to keep up with the predictable demand and the service consistently falls short in responding to calls, particularly at busy times
- Clinicians work inconsistently so that, depending who is carrying out the telephone assessment, the service will sometimes fail to meet the standards
- The service has not been measuring its performance in line with the standard so is not focusing on improving performance, believing that they comply or are close to compliance

- The level of urgent cases is so high that it becomes more difficult to respond to so many within 20 minutes

Commissioners and providers should not make the mistake of thinking that the standard is unachievable. It is possible to fully comply with the standard, but the underlying cause of the shortfall needs to be understood and addressed.

We provided information also about the percentage of less urgent cases that are definitively assessed in 60 minutes and the percentage of all cases assessed in 20 minutes. In general, services that have properly matched capacity to predictable demand begin definitive assessment of at least 60% of all cases in 20 minutes. The measure of time to definitive assessment is not just important in improving the patient experience (though, as described above, patients see responsiveness as closely linked to the quality of care). If the delay is too long, then the chances of the concerned patient or carer making their own decision to go to hospital is increased, with consequent resource implications for the local health economy as a whole, and secondary care in particular. In addition, in services meeting the standards, the telephone assessment of the vast majority of all cases will take place within 30 minutes providing a valuable safety net in case any potentially life-threatening or urgent cases were to have been missed by the non-clinical call-handlers.

QR12 Time to face to face consultation

The quality requirement recognises three priorities after definitive clinical assessment: emergency, urgent and less urgent, with the target time within which the consultation should begin being 1 hour, 2 hours and 6 hours from the end of the definitive assessment. For the benchmark comparison we looked at the percentage for urgent cases and the percentage seen within 2 hours. Performance against this standard was generally considerably better than against QR9 and 10, with over half fully compliant with the standard and 80% partially compliant.

Figure 9 showing the range of performance in seeing urgent cases face to face, with the majority compliant or partially compliant with the standard
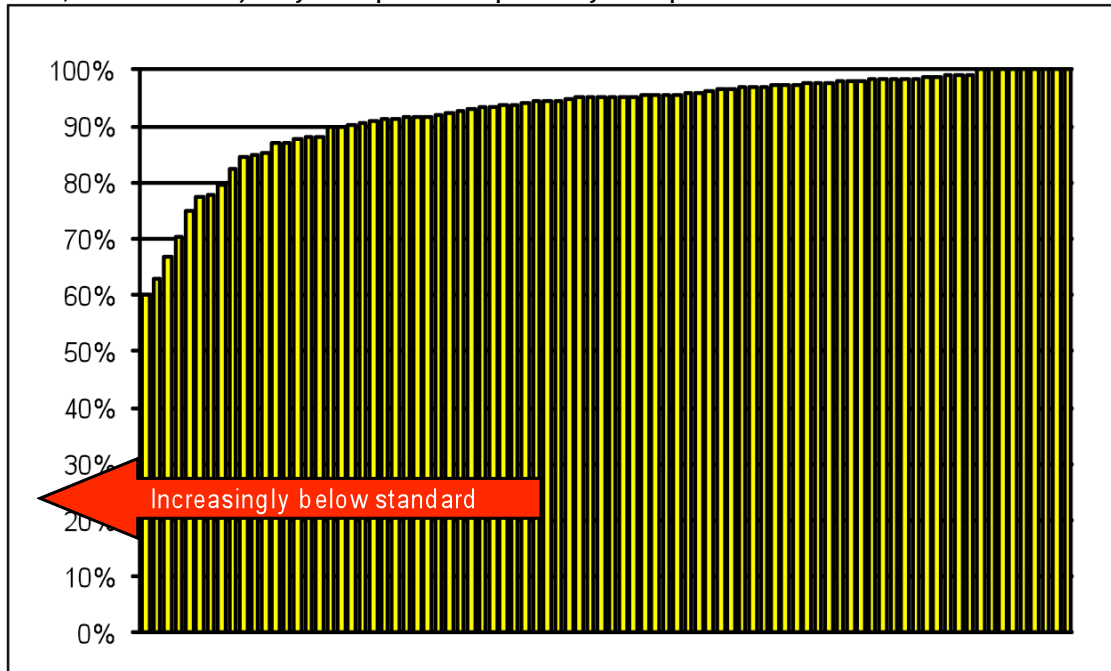
Again there were very wide variations in the percentage of cases identified as urgent or emergency, with some services identifying no emergency cases in the four week sample. Whilst we recognise that clinicians will have called an ambulance if there are indications of an immediately life-threatening condition, we are sceptical about such services. Among several thousand cases it seems to us that one might expect there to be at least a small number of cases that the clinician felt should best be seen face to face within an hour, for example to provide pain relief to a palliative care patient or if a patient was refusing to allow an ambulance to be called.

Reporting variations between services

There are a number of cases where provider services do not report in line with national standards. Whilst we have no issue with individual PCTs and providers agreeing to measure performance in a slightly different way, we thought it helpful to highlight some of the differences that we have encountered. The software system supplier can advise how best to configure the system so that reports are in line with the national quality requirements. However, we would recommend that if any change is made, some reports are run for a parallel period so that the effect of the

change can be understood and recognised or adjusted for when looking at any historical trends.

Inadequate systems or lack of integration

In some cases we found services whose system was unable to report against the majority of the national quality requirements. We were told that this had occurred because the commissioning PCT had imposed a system on the provider. Equally, we came across cases where investment had not been made in the system, or the different elements of the system were incompatible so that services were operating 'on paper' and the clinical notes were not entered on the system or the type of consultation was not identifiable so that accurate reporting against the national standards was impossible.

Whatever the rights and wrongs of how the situation had arisen, in our view it is unacceptable for any service to operate without:

- Being able to measure performance and demonstrate whether it is in line with the national quality requirements
- Having available a suite of reports that allows the service to look at the performance of individual call-handlers and clinicians, the clinical decisions made, the dispositions, their productivity etc. so that the service can identify outliers and drive up clinical and operational performance through a proper governance process that feeds comparative information back to the staff involved. Emphasis on these aspects was particularly highlighted in the recent CQC interim findings that made clear the responsibility of PCTs to *"look at details like the quality of clinical decisions, the efficiency of call handling, the adequacy of staffing"*..
- Having communication links so that details are transferred electronically from NHS Direct when they refer a patient to the out of hours service. These had been in place in all PCT areas but it appears that with some new contracts and systems the communication links are no longer in place
- Being compliant with the directions issued by the minister after the Penny Campbell case *[Primary medical services (out of hours services) Directions 2006]* that the fully recorded clinical notes of earlier consultations should be available to a clinician that may see or speak to the patient later whether they are within the out of hours service or within the patients' practice on the next working day

Whilst we are unsure about the legal liability of the PCTs and providers, we are absolutely certain that the organisations and individuals involved will be exposed to

severe criticism were there to be a serious incident that arose from some of the gaping holes in the arrangements that appear to exist in a few places.

Counting of cases

There are variations or inaccuracies in something as apparently simple as the counting of cases (something that is usually apparent from comparison of the reported number of cases with the data extract). Sometimes this is caused by coding inaccuracies that are described below. However we have also come across a number of reasons for significant differences such as:

- Some services count the number of consultations or contacts rather than the number of cases. In common with most services we define a case as being the whole episode of care from the initial contact with the service through to the case being closed by the service in the expectation that the patient will have no further need to contact the out of hours service. Since a case may often involve both a telephone consultation and a face to face consultation (and sometimes more), it is clear that a count of consultations or contacts will produce a higher number than a count of cases. Without clarity on this, the provider and commissioner cannot compare even the most basic of information about the number of cases per 1000 of population or the cost per case.

- Some services filter their reports in such a way that they systematically exclude or miscount the number of cases and omit these cases from the measurement of performance. This can be a problem particularly when a number of services share a hub. The most common example that we have come across is services that use the doctors' operating group to distinguish the patients of one PCT from another, or one provider from another. Unless steps are taken to allocate patients who are visiting, patients who are unregistered or who cannot remember the details of their own doctor to one or other of the different doctors operating groups, these cases will be excluded from all of the reports.

- Providers have set up reports that were intended (for example) to separate the out of hours cases by referring to a cover rota. Over time, however, the importance of maintaining this cover-rota to ensure that reports are complete has been forgotten so that new practices have not been set up on the cover rota.

- All services have some cases that are cancelled. Generally these are cases that are subsequently excluded from the analysis because, for example, they duplicate another case that was already within the system and being dealt with. However some users are in the habit of cancelling cases that were valid out of hours cases but which required no further action from the service, for

example because the patient went to or was referred to the
hospital. Such services will often exclude these cases from their reports when
they run them. By mixing some cases that they certainly would not normally
want to count as a separate case (such as the duplication of a case) with cases
that they certainly should report (such as a case where the patient was
referred to the hospital), providers get in a muddle and appear to under-
report these important cases for this reason.

The importance of coding structures and consistent use of the system

Many systems have grown over a period of time in which the software has been
developed and considerable additional functionality and features have been added.
Unfortunately the coding structure used by many services has not been updated and
many show signs of the sort of neglect that is typical of systems of a certain age. The
coding conventions have become even more confused where a hub is shared and
different services have not worked together to ensure consistency, simplicity and
rigour in setting up the options.

This summary is not intended as a criticism of any particular service. The process is so
well recognised that it could be described as a natural law of entropy in that the
disorder in a system will naturally increase over time. But, unlike in thermodynamics,
energy is required to re-instate order in coding. And it is worth it. Without good
coding structures it becomes very difficult or impossible to report accurately, and the
multitude of really useful system reports cannot be filtered meaningfully to provide
good management information. With good coding structures, reliable information
can be produced on all or part of the service using standard reports.

Our experience of looking at Adastra systems has convinced us that it would be
worth the vast majority of services rationalising the codes that they use and
exploiting the newer fields and functionality that are now available. Whilst we have
provided in the appendix considerable detail about specific codes and how they are
used and abused we have confined ourselves here to a short list of principles to be
applied in defining the coding structure.

- Make sure that you use one code to do one thing and one thing only – you can
  sub-divide logical groups but should avoid mixing two completely different
  things together. If the code is intended to allow you to describe the priority
  then it should be used for that and that alone, not as an alternative or
  additional way of identifying the disposition or case type.
- Make sure that the options you offer the user are mutually exclusive and
  collectively exhaustive. Faced with picking a code the user should always be
  able to select one that is appropriate. They should not find that there is no

code that fits that case, or that there are two or more possible
codes that are relevant.

- Codes should be unambiguous. If you have classified cases as, for example,
  District Nurse case type, is this because the call has been taken on behalf of
  the district nursing service, or is it because the call that came into the out of
  hours service has been passed to the district nursing service for them to visit
  the patient?In some services it sometimes seems to be used for both, making
  it impossible to identify one type of case from the other.
- The number of codes should be short enough that the decision is easy and the
  user does not have to search extended drop-down lists
- Codes should, as far as possible, be readily recognised and understood by the
  users
- Codes should allow you to report as you need to and to sub-divide and filter so
  that you can segment the activity in a way that is useful and reflects the
  responsibilities within the organisation.

In the appendix you will find considerable detail about some of the uses and abuses
of coding that we have come across, together with some details of some of the fields
that are available and seem to us to be under-utilised (such as the service field which
not only makes the reporting of different services much easier but allows for
different workflows and choices depending on the service that has been selected).

Services should also recognise that all users will need to be reminded about correct
use of the system and codes. This is much more than an occasional standard email. If
providers want to ensure that coding is accurate and will support sophisticated
analysis because it has been consistently applied, then they should recognise that it is
necessary to regularly review codes and feed back to individuals promptly if any
inconsistency is spotted or if the system has been misused in any way.

QR2 Reporting details of consultations to the patient's practice

Nearly all services have their systems configured in such a way that details of any
contact is automatically transmitted to their GP practice, increasingly electronically
but in some cases by fax. The transmissions automatically run a number of times
through the out of hours period and the times are chosen such that by eight in the
morning very few (often just the details of any consultation still in progress) remain
to be sent. Because of this (barring system and communication failure) there are
very few cases that have not been sent by 08.00 as required by the quality
requirement. Because of this near-universal compliance we do not devote time in the
benchmark reports to a detailed comparison of performance against this standard,
although we do report performance and highlight any under-performance for
investigation.

There are, however, differences in the way that the figures are reported. Some services report the percentage sent (in line with the standard) whilst others measure the percentage where an acknowledgement of receipt has been returned. In the latter case, though it is no fault of the provider service, the report will identify occasions when, for example, the fax was not transmitted because the machine within the practice had not been turned on.

QR8 Telephony

There is a great deal of inconsistency in reporting against this standard. The standard and guidance is quite clear. Where the service has an initial message for the caller (the vast majority do) the requirements are:

- That the message should be no longer than 30 seconds
- That the percentage of calls that receive an engaged tone should be less than 0.5%
- That the percentage of calls that abandon later than 30 seconds after the end of the message should be no higher than 5%
- That all calls should be answered within 60 seconds of the end of the message (full compliance is achieved if the service does better than 95% on this measure)

In practice most telephony systems will hold so many calls in a queue that the caller will never receive an engaged call. There may, however, be instances when calls are automatically transferred from a practice when the caller will be unable to get through because of the restricted number of lines available at the practice. So far as we know there is no practical way of recording these failures so our recommendation is that before call-forwarding is instigated the appropriate number of lines to ensure compliance with the standard is calculated using the Erlang formula.

Unfortunately many services do not collect the information to allow them to report in line with the quality requirement. There are a number of reasons for this:

- No equipment is installed to measure the above. Often this is an initial problem when the service is moved, for example to a hospital site where there may be complexities because of the hospital phone system or may simply be because of a short delay in installing the necessary equipment. However there are instances where the problem appears to be long-standing
- The equipment or software does not allow all of the information required to be captured, for example because the time fences cannot be set at 30

seconds and 60 seconds after the end of the message or the
reports do not allow the relevant information to be extracted

- The PCT and provider have not understood the standard or the guidance so
that they fail to collect the relevant details. One specific point that many
services have missed is that they should not count calls that are abandoned
during or within 30 seconds of the end of the message in either the total of
abandoned calls or in the total of calls used as the divisor. *[This long-standing
definition of abandoned calls that can be discounted was confirmed in the
definition provided as part of the "Commentary on the National Out-of-Hours
Quality Requirements and their Performance Management" issued by the
Department of Health in October 2004]*

- The system is not set up to count calls separately during the out of hours
period or covers lines used for non-clinical incoming calls so that the measure
includes other calls that are not relevant to the standard. Some services get
round this by running numerous reports for each period from 18.30 to 08.00
and then assembling the data in a spreadsheet - but we would recommend
that if the volume of other calls included is likely to significantly affect the
measure of performance that services separate incoming calls in such a way as
to simplify reporting of this measure.

So few services provided information that inspired confidence that they were
measuring in line with the quality requirements that we have chosen not to compare
services, but simply to feed back the analysis based on the information provided.
Within the appendix we have provided a simplified worked example of the
calculation method that we use.

QR9 Identification of urgent cases.

We have already described above the wide variation in the proportion of cases that
are identified as urgent. We do not intend to repeat this, or highlight again the
dangers if the proportion is either too high or too low. However there are some
specific differences in the way that such cases are identified and we thought it
helpful to highlight these.

Before identifying some of the differences it is perhaps worth remembering that in
every out of hours provider far greater attention is focused on this area and much
more effort is put into the systems and training than is typical in any other area of
primary care. When we produced the report into "Urgent Care - a practical guide to
transforming same-day care in general practice" we found that in most practices
much less attention was paid to ensuring that receptionists would reliably recognise
a potentially urgent case.

Out of hours services use two main approaches to support call-handlers in identifying life-threatening and urgent cases - paper based protocols or, increasingly, clinical decision support software. Whichever is used we believe that it is at least as important to train call-handlers so that they:

- Recognise the symptoms of any condition that is potentially life-threatening
- Recognise the triggers that might indicate that a condition is more urgent
- Understand how to ask relevant questions without prompting the caller to respond in a particular way
- Take account of the circumstances, age and any special notes about the patient
- Understand that they will need to make a judgment, because in many cases the information available will be limited
- Recognise the need to err on the side of safety
- Have developed approaches to cross check information from the patient when necessary

The role of the call-handler is an important one. The skills outlined above are not easily learnt. Unless a service routinely reviews and compares the priorities set by each call-handler during their training period and regularly thereafter, feeding back information and reviewing cases with the individual by listening to recordings, then there will be very wide variation between individuals. Whilst there will always be some dofference between indviduals the implication, if this is too large, is that some call-handlers are over-prioritising whilst others may be missing urgent cases. Neither is good for patient safety or for the consistent operation of the service.

In addition to the differences that arise because of the factors listed above we have also observed that the percentage of urgent cases often rises when the provider starts to use clinical decision support software. Some of this change may come from the focus on the area that is an inevitable part of introducing such a system and for many other reasons, but there are certainly instances where the increase is because the paper protocols previously used did not recognise as many potentially urgent cases.

We would suggest that anyone looking to review their protocols and training might want to check that they have properly considered each of the areas listed below (this list is drawn from the Adastra prioritisation protocol, but a similar range of topics is covered by other decision support systems and they include detailed questions or algorithms below each heading that provides the user with guidance on disposition and priority):

- Patient unconscious
- Pregnancy

- Rash
- Allergy
- Fitting
- Chest pain
- Breathing
- Injury
- Suicidal

- Burns Scalds
- Stroke TIA
- Death
- Terminally ill
- Child under two
- Other

Many paper-based protocols and much training for call-handlers fail to cover such a wide range of symptoms and situations

QR9 mapping of priority from NHS Direct

Nearly all providers have calls transferred to them from NHS Direct. These patients have dialled 0845 4647 and have generally been assessed by a nurse who has identified that it is appropriate for them to receive care from the out of hours service. Typically about 10% of an out of hours service calls are transferred in this way. In some areas, NHS Direct also provide the front-end of the out of hours service, typically carrying out the initial call-handling and some or all of the clinical assessment.

Through the benchmarking exercise we have identified that there is a variation in the way that NHS Direct codes have been mapped into different provider systems so that the same code can be treated as urgent by one provider and less urgent in another. These differences, together with differences in the process adopted mean that some services have well over 50% of cases that they receive from NHS Direct classified as urgent whilst others have levels nearer to 10%. NHS Direct use one clinical decision support system and they monitor users closely to ensure that outcomes are consistent and as safe as they reasonably can be.

Because of this difference we have recommended that NHS Direct review and clarify which of their outcome codes should be classified as urgent and which as less urgent as defined by the out of hours quality standards and ensure that this is made crystal clear to all services so that the mapping is consistent. As the experts in their system and with the knowledge of which algorithms lead to which dispositions this responsibility seems to lie with NHS Direct and is not something that we believe that individual out of hours services should vary. Until this is done, NHS Direct, the PCT and the provider service seem to us to be exposed were there to be a serious incident with a case that was defined as less urgent even though the same case and disposition would have been identified as urgent elsewhere.

QR9 time to definitive assessment

There are a number of differences over which point is taken as being the start and end-point for measurement against this quality requirement. The requirements themselves specify that the measurement should be from "the call being answered by a person" to the start of the definitive assessment. The definitive assessment is the last one if there is more than one assessment and which is *"In practice... the assessment which will result either in reassurance and advice, or in a face-to-face consultation"*

Many providers do not measure performance in line with the definition of the quality requirement.

- Many services measure this from the end of the initial call from the patient. The standard measures from the start of the call.
- Many services measure to the initial clinical assessment not to the definitive assessment. Services that have a split service with two providers, one responsible for call-handling and some clinical assessment and one providing face-to-face care, need to look particularly carefully if they are to make sure that they report this accurately. In some cases where the two providers operate on different systems and the information is not passed across it seems to be so difficult to assemble the information to allow proper reporting against the standard that it is never done.

As far as we are aware, except where there is a split service and the information is not carried across from one system to another, all services report against this standard (as they should) using the priority on reception defined by the call-handler.

There is one minor difference that occurs in the way that some services account for patient attributable delays. These occur typically when the clinician tries to ring back the patient (sometimes several times) within the specified period (20 minutes for urgent and 60 for less urgent) but fails to get through because the phone is busy or is not answered. In these cases the current Adastra system allows a report to be run that will exclude such cases from the calculation provided that the clinician identified the failed attempt properly on the system and provided that the attempt was made within the prescribed time depending on priority.

In this instance we quite understand why a PCT or provider may choose to monitor their performance after discounting such cases - the argument goes that if the failure is no fault of the provider why should it count against it (though it should be remembered that if the call had been warm-transferred patient attributable delays of this type would not occur). However, since many services do not or cannot identify that the delay was attributable to the patient we have, as described above, taken no

account of patient attributable delays in the main measure that we use for comparison. However, where we can, we report what the performance would have been if all calls had been definitively assessed at the first attempt. Providers and PCTs might like to note that the typical difference between the two figures is 2 or 3% because of patient attributable delays. Where the difference is larger than this it is likely that it is caused by the double-assessment that we described above.

QR10 Time to definitive assessment for walk-in patients

Most out of hours services have only a small number of walk-in patients and the clinician that first sees the patient will complete the assessment as part of the full consultation. However some services, typically where the service is integrated with another walk-in service such as a minor injuries or walk-in centre, adopt a different process.

In these services patients are 'triaged' soon after arrival by a clinician. This step is the face to face equivalent of the telephone assessment designed to identify any life-threatening emergency and/or to assign a priority to the patient and therefore to identify within what period they should receive a full consultation. After this initial assessment the patient goes back to a waiting room before being seen later for their full consultation.

For services that operate in this way to report against QR10 (time to definitive assessment) or QR12 (time to face to face consultation) they have to distinguish between the two different types of face to face consultation. In our experience few integrated services do this so it becomes impossible to report against this standard. Whilst the service could argue that such patients fall outside the out of hours standard if they have walked in to the minor injury or walk in service (so will typically only be assessed against the 4 hour target for completion of the case), we feel that an opportunity is being missed to report on an aspect that is important. In addition where a patient has telephoned the service (so should be counted in the analysis of out of hours performance) and is reassessed on arrival it is essential for accurate reporting to be able to separate out the assessment from the full face to face consultation.

QR12 Time to face to face assessment

The quality requirement makes clear that the time is measured from *"after the definitive clinical assessment has been completed"* to the start of the face to face consultation.

As described above many services do not properly identify the definitive clinical assessment and we suspect that this may mean that they do not report accurately against this standard either. However there are some other variations that we have come across in reporting against this standard.

- Some services discount any cases where the delay was attributable to the patient.
- Some services only measure their performance against this standard on home visits, arguing that otherwise they are dependent on the patient as to whether they are seen within the target time.
- Some services use the priority on completion to evaluate performance.
- As we have described above when looking at double assessment and split services, some services allow (or even encourage) clinicians to telephone a patient when the case has been passed across to them as a face to face consultation to take place in the home or at the centre. The case is then recorded incorrectly and reported against as the home visit or attendance at the centre.

None of the above seem to us to be in line with the national standards and the intent behind them for the following reasons:

- If a clinician has identified a case as an emergency or urgent case then we believe that it is the responsibility of that clinician to make sure that the patient or carer recognises the importance of an early face to face consultation. For this reason we expect the conscientious clinician carrying out the telephone assessment should make sure that the patient will be seen within the appropriate timescale by checking that the importance is understood and that transport arrangements will permit the patient to arrive within the appropriate timescale. Because we see that responsibility for checking this rests with the service, we do not believe that patient attributable delays should be discounted in calculating performance against this standard.
- Providers should measure performance against this standard using the priority set when the case was clinically assessed (or if it was not entered by the assessing clinician, the priority set on receipt which the assessing clinician presumably accepted was appropriate). Not only would any reporting that looks at priority on completion raise questions about whether such priority, set with the benefit of hindsight both about the patient's real condition when seen face to face but also about the actual time to the consultation, would be as objective but the guidance to the standard clearly indicates that *"Definitions of emergency, urgent or non-urgent will be made by the clinician who completes the definitive clinical assessment of the patient's needs and the*

> *record system will need a means of capturing and auditing this information."* Some services use the priority on completion which is not in line with the standard.

- No reliability can be placed on the reports from any provider where cases that were actually dealt with on the phone are recorded on the system as face to face encounters.

Lessons for commissioners and providers

Matching capacity to demand

Where services are consistently falling short of the standards for timeliness (time to definitive assessment and time to the face to face consultation), providers should investigate and report to commissioners how performance varies by time of day for both weekdays and weekends. They should identify whether the sub-standard performance is associated with an inconsistent response from individual clinicians or because there are insufficient staff available to meet predictable demand. Commissioners should expect to see realistic plans for ensuring that performance is consistently at least partially compliant with the main standards.

Ensuring robustness and clinical safety when calls are received

Providers are expected to have a *"a robust system for identifying all immediate life threatening conditions"* and a demonstrably *"clinically safe and effective system for prioritising patients".* We think that this capability should be demonstrated to the commissioner so that it is the responsibility of those with clinical and process understanding from the commissioning side to satisfy themselves that the process is robust and fit for purpose. We would also expect the overall process to recognise the different needs of the very young, the very old and to take account of any special notes about the patient.

Commissioners should look both at the training and the protocols or decision support software. They should also ensure that there is room for the call handler, on occasion, to use their judgment to prioritise cases even when they fall outside the normal guidelines. Since it is impossible to define every possible combination of circumstances that might lead to the case needing to be prioritised, the training should aim not just to equip call-handlers to recognise a narrow range of conditions, but also enable them to take a wider view and to spot combinations of circumstances that might be urgent.

When looking at the protocols or decision support software, commissioners should ensure that it covers a sufficiently wide range of situations that there is little danger of a potentially urgent case being missed. The list of areas identified as a default in the standard Adastra prioritisation software may help with this (it should be noted that individual services can configure the tool within Adastra to operate with different topic headings, questions and rules).

Providers should include, as part of their governance processes, a review of the nature of cases where priority is escalated during the patient's journey so that they can identify whether escalation occurred because the potential urgency was missed on receipt and can identify any lessons for individuals, or spot if the process, algorithms or protocols need to be refined and if any points need clarifying for users and should be reflected in future training. Commissioners should expect to see evidence of such a governance process.

QR2 Reporting details of consultations to the patient's practice

Commissioners should check that provider systems are configured in such a way that information about consultations is automatically sent to GP practices by 08.00 on the next working day, and that the small number of cases that finish close to this time are sent shortly afterwards. Where the system does not support such details being sent automatically, the commissioner should make sure that the process used meets the standard and monitor performance in sending the information more closely.

Commissioning PCTs might also like to ask the provider for periodic reports on failures where GP practice equipment or systems were not able to receive the details by 08.00,so that where such problems occur regularly they can be investigated.

Commissioners should also satisfy themselves that (as required by the directions issued in 2006) "arrangements are adequate to ensure that clinical notes of consultations are fully recorded and subsequently accessible" both within the service and to the patient's practice. Commissioners should satisfy themselves that:

- The details of all relevant consultations are captured on the system in a timely fashion so that they are available when required by another clinician. Particular attention should be paid to any instances where details are recorded on paper to ensure that relevant information is copied onto the system quickly.
- The notes, whilst being concise as they reasonably can, provide another clinician with a clear picture of the history taken, the findings from any examination, the results of any tests, the diagnosis made (if any), any

medicines prescribed, any treatment given and any instructions
given to the patient or for another clinician about the need for any follow-up.
- The provider has in place adequate audit and feedback processes to ensure
  that such information is consistently collected and recorded regardless of
  which clinician was involved and their location.

QR 8 Telephone standards

Where practices, providers or commissioners decide that calls should be forwarded
automatically from the practice number to the out of hours service, commissioners
and providers should ensure that the arrangements are such and availability of lines
is sufficient that patients will not receive an engaged tone more frequently than in
0.5% of the times that they ring even at times that are typically busier. Use of the
Erlang formula is one way that this might be checked.

Providers and commissioners should check that they are collecting information to
allow reporting to be in line with the standards and guidance. Whilst we have tried
to make sure that the description above and the example in the appendix makes
clear what information is required, we would be happy to respond to any queries
about this from providers and PCTs that are part of the benchmark so that we can
check whether the data provided can be compared with the standard.

In the future, to simplify reporting and in the setting up telephony systems, we
suggest that (when they next review the standards that apply to out of hours
services) the Department of Health consider aligning the period within which
abandoned calls may be discounted either with the end of the message or with the
time within which all calls should be answered.

QR 9 and 10, definition of urgency

In addition to the checks on the process by which urgent cases are defined described
above, commissioners should use the information from the benchmark to check that
the level of urgent cases is not so far from the norm as to cast doubt on the
robustness and clinical effectiveness of the process. Should the commissioner not be
satisfied that the service meets the necessary requirements in this respect, they
should insist on the service beginning definitive assessment of all calls within 20
minutes.

NHS Direct should clearly define the out of hours priority for each of their outcome
codes and providers should make sure that priorities are mapped accordingly.

QR 9 and 10, measurement of time to definitive assessment.

Providers should report on the time to definitive assessment in line with the standard, counting from the time the call was answered by a person and measuring to the start of the definitive assessment (the final assessment if there is more than one). Where the service is split between two providers the commissioner should make sure that reporting is in line with the standard and that there is no 'time gap' in the reporting between the two services.

Where patients walk in to the service and receive an initial 'triage' (a clinical assessment that is intended to identify the priority but after which the patient has to wait for a full face to face consultation), the service should ensure that this is clearly distinguished from a full consultation and that reports correctly recognise the difference between them.

QR12, time to face to face consultation

Providers should ensure that their reporting is based on the priority at the time of clinical assessment (or on receipt if it has not been changed by the clinician conducting the assessment) and not based on the priority on completion.

Providers should ensure that no clinician is abusing the system in such a way that a telephone consultation is recorded as a face to face consultation, thereby undermining the ability to report against the quality standards and misrepresenting the disposition.

Providers should report against this quality requirement without adjusting for patient attributable delays and recognising their duty of care to make sure that the patient understands the timescales within which they should be seen. They are of course free to provide additional information that identifies what proportion of failures against the standard can be attributed to the patient.

Quality of governance processes

The benchmark identifies that there are significant differences between services. But these are much smaller than the differences between individual clinicians operating within a service.

This section aims to look at the sort of differences that are common and to suggest how the governance processes that some providers are using or developing might ensure that the service operates in a consistent way. This is probably one of the key aspects that the CQC highlighted in their recent interim findings when they referred to the need for PCTs to look at *"the finer detail of the actual care patients receive, to ensure the service is safe and meeting people's needs."*

Variation within a typical service

We have chosen to illustrate the variability that frequently exists between individual clinicians within one service. The graph below looks at the proportion of cases given telephone advice by each individual clinician. The data is from a six month period, excludes those clinicians that assessed comparatively few patients and is from a service that has a doctor only model. There is a considerable difference between individuals. Depending on which clinician answers the call the chances of receiving telephone advice vary from less than 30 % (for the 15 clinicians on the left) to over 60% (for the 19 clinicians on the right).

In the second graph below we look at the length of the telephone consultation, but we focused only on those cases where the clinician decided that the patient should be seen at the centre. It is suggested that cases that end up requiring a home visit are often more complex so may take longer to assess and that it is often worth investing a little more time to be sure that the telephone advice is clearly understood. By looking only at those cases where the patient was invited to the centre there was an expectation that the call would not be extended for too long, because the clinician should recognise the need for the patient to be seen face to face reasonably quickly. In reality there is a big difference here too, with some clinicians completing fewer than 20 % of cases in four minutes (the 20 on the left) and some over 75% (the 15 on the right).

Figure 10 showing the typical variation in dispositions between individual clinicians within a service

Percentage of calls converted to advice - only including those that handled more then 25 advice calls
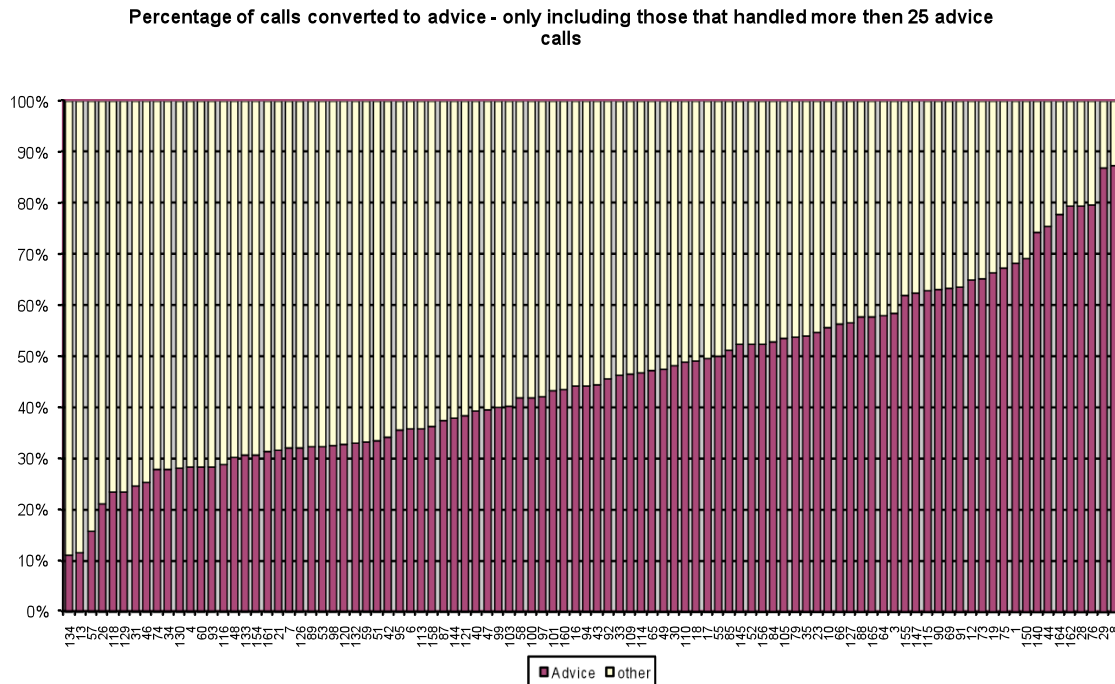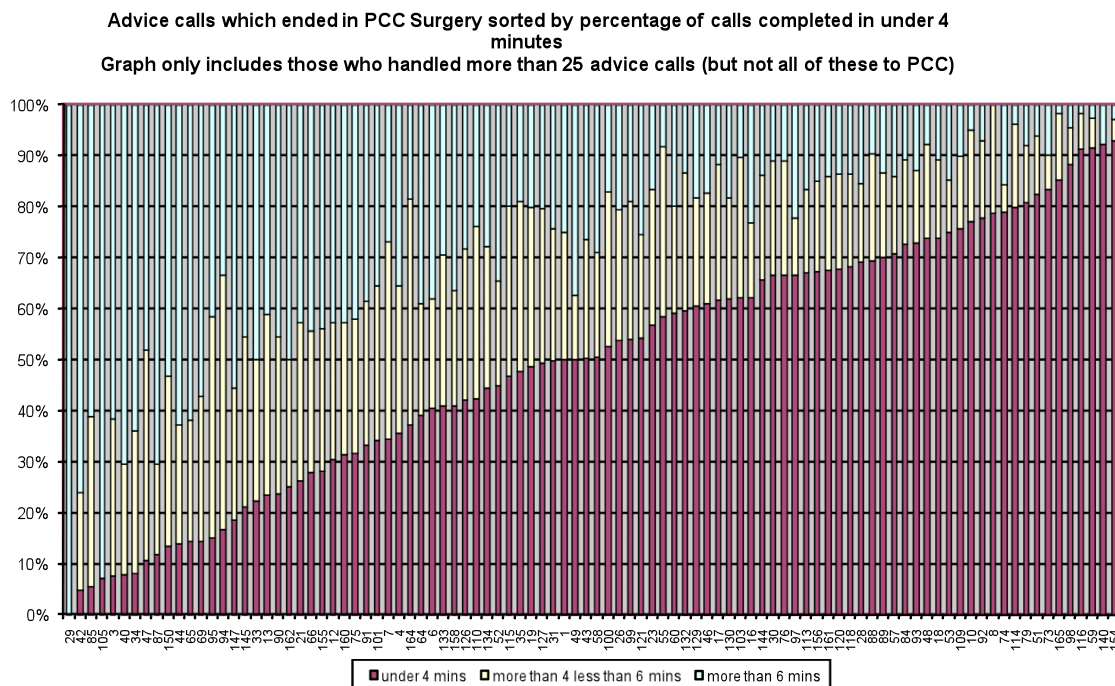

Figure 11 showing the variation in the length of the telephone assessment even when only looking at cases where the conclusion was that the patient should be seen face to face

Advice calls which ended in PCC Surgery sorted by percentage of calls completed in under 4 minutes
Graph only includes those who handled more than 25 advice calls (but not all of these to PCC)

In the service above, although there were some cases where clinicians
leant towards spending more time on the phone with the patient and were thus able
to complete more cases as telephone advice, there were also some who were
comparatively slow and cautious in the proportion that they felt able to complete as
telephone advice. Equally there were some clinicians where the rapidity of the
assessment and the proportion of telephone advice raised questions about whether
the telephone consultation was long enough for a full assessment.

We have no reason to believe that this sort of variation is not typical of many
services. We have also looked at the percentage of cases that call-handlers complete
within two minutes, the percentage of cases identified as urgent or potentially life-
threatening and found similar variation between individuals.

The reason for highlighting this is to demonstrate how important it is for a service to
understand this variability, to manage the service with an understanding of the
differences between individuals and to work towards reducing the range of variation
in this and other aspects if the service is to operate in a consistent way and reliably
meet the quality requirements.

The governance framework

Whilst the examples above provide an indication of why it is important that
governance processes look at some of the detail of the operation, they also give an
indication of one of the key elements that underpin a strong governance framework.
Staff should see how they compare with their peers and understand what they need
to change to improve the service offered to patients.

We would expect that feedback to clinicians would, over a period, include the
following sorts of information allowing individuals to understand how they compared
with others:

- Feedback on timekeeping issues and any failure to work as expected during
  their paid hours
- Feedback based on an analysis of a sample of calls over the record keeping
  and use of the system - typically we would expect that this might be assessed
  using the Royal College of GPs tool-kit or something similar
- Feedback based on listening to a small number of recordings of consultations
  again making use of an assessment criteria

- Feedback of the proportion of patients given telephone advice, a home visit or invited to attend the care centre based on cases over a considerable period of time (perhaps 3 or 6 months)
- Feedback showing the comparative numbers of patients referred towards hospital
- Feedback from any analysis of consistency of coding or use of the system
- Feedback on productivity, looking both for clinicians that may not be devoting enough time to patients and for those that do not pull their weight
- Feedback from their peers and from other staff about their contribution to the service and any concerns or issues that have arisen from specific incidents
- Comparison of common or sentinel conditions across clinicians looking at the consistency of clinical coding, whether NICE guidance was followed and the appropriateness of any prescribing or treatment recommended
- An evaluation of prescribing across clinicians to spot any with a tendency to over or under prescribe.

In collecting this information to be shared with clinical staff, services should recognise that there will be variation that occurs for any number of reasons, particularly the different case mix that may occur at different times of the night and the variability in the volume of cases, but also because of the way that some teams may choose to work. The intention is that in most cases the discussion will be an open one that will allow the clinician to reflect on his or her clinical practice and consider if there are any areas that might be worth focusing attention on. There may, however, be some outliers where the performance is such that specific focus is necessary and, in some cases, this may lead to the service deciding that they no longer wish to use a particular clinician. Even where there is no question about their clinical competence, the service may find that it does not wish to try to accommodate an individual who, for example, compromises the effectiveness of the service because of their slowness at telephone assessment. This will be rare, but what is perhaps more likely is that the strengths of individuals in one area and their discomfort in another will be explicitly recognised and can be factored into the planning of the service, making sure that it operates more effectively as a result.

Regrettably many services fail to provide this sort of feedback to individual clinicians (or to provide similar information and support to call-handlers and others within the service). It is perhaps for this reason that the CQC particularly highlighted the need for the NHS to improve the monitoring of out of hours services, the need for PCTs to ensure that they look at details like the quality of clinical decisions, the efficiency of call handling, the adequacy of staffing and doctors' training in order to rigorously monitor performance.

Lessons for providers and commissioners

The out of hours sector is rich in data - it would be negligent not to use this information to compare individuals and to help them improve the service that they offer - to the benefit of patients but also to improve satisfaction among the staff from a job that they can thus do better.

Governance is not just about identifying aberrant clinical or non-clinical staff – it is about *"continuously improving the quality of services and safeguarding high standards of care".*

Making use of the questions attached to the benchmark report

As part of each benchmark report we attached a list of questions about governance. We urged commissioners to use this as atopic guide to allow them more fully to understand how the service operates and the extent and robustness of governance processes . Although we will be developing the list further (and we will make it available in the spring of 2010), we commend it again to both providers and commissioners.

Comparing individual staff and clinicians on key measures

We have illustrated the variation that is typical between clinical staff, and we know that in many instances there are similar variations between non-clinical staff with, for example, some identifying almost no urgent cases and others accounting for a very large proportion of them. Identifying this difference and then moving on to look at prescribing or treatment of common or sentinel conditions seems to us to demonstrate that the service has become a truly competent provider of healthcare that is properly making sure that the service is consistent, safe, timely and cost-effective. Crucially we believe that this information must be shared with all staff if the service is to be an example of the *"self-improving NHS".*

What does a good service cost?

What does good look like?

After the first round of the benchmark a number of PCTs and providers asked us to cast more light on the wide range of cost that services seem to be provided for. They wanted to understand whether a good service could be provided at a reasonable cost or whether only the more expensive services provided a good service. This begs the question 'What does good look like?'.

We are cautious about narrowly defining this, and very cautious about scoring services when they operate in different environments, sometimes with excellent and sometimes less good in hours and out of hours services.

We have chosen to exclude services on a number of criteria so as to arrive at what we have chosen to call 'good all-rounders' - though strictly they might more accurately be defined as those services that did not fail any of the criteria that we chose.

Our criteria were:

- The service should not have less than 5% of urgent cases on receipt. We have concerns that such services might be missing some potentially urgent cases so felt that these should be excluded
- The service should be in the upper half in a composite measure of the helpfulness of the health professionals (above 67% score)
- The service should average better than 90% on the main measures of timeliness (definitively assess urgent cases in 20 minutes, less urgent in 60 minutes and all in 60 minutes and seeing patients face to face urgent in two hours and less urgent in 6 hours).

We could have chosen different criteria, we could have weighted some aspects more heavily than others, but we felt that these criteria excluded services on a reasonable basis whilst leaving a mix of services of different types.

We chose to identify these services by circling the data points on the various scatter graphs that follow. Before we look at cost, though we felt it useful to include the same graph that we showed above (figure 3) but with the data-points of the 'good all-rounders' circled. The purpose is to emphasise that we have chosen those that scored more highly on the patients' perception of the helpfulness of the health

professional (so it is no surprise that they also score well on care received),and we have also chosen those that were more responsive on a mix of measures (so it is no surprise that they also generally began definitive assessment of most cases in 20 minutes.

Figure 12is the same as figure 3 but demonstrates that the 'good all-round' services which have been identified are generally responsive and seen by patients as providing good care
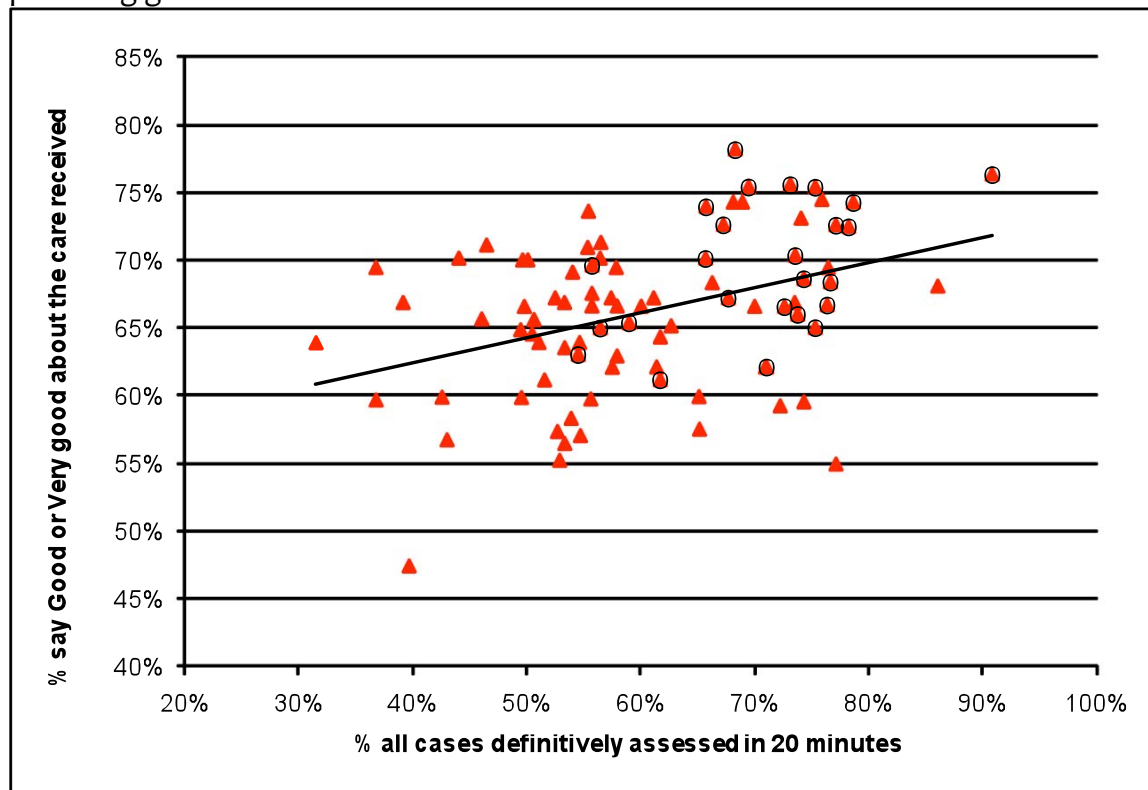
Cost per head and cost per case

We look at the cost per head of the out of hours cases. In making the comparison we have chosen to plot against population density and have defined three broad bands - rural, mixed rural/urban and city/urban. The geography and population density significantly affects the operating cost of a service, with those in rural areas needing more centres to be open (although they are lightly used) and requiring more home visiting clinicians to cover much greater distances between visits. It is no surprise, therefore, to find that cost generally increase as one moves towards the more rural services.

What is surprising is the wide variation in cost per head for services that cover areas that are essentially similar in respect of their population density. However the range is much narrower and the indications are that the 'good all-rounders' are nearer to the middle range of cost, with the majority between £6 per head and £11 per head.

Figure 13 Showing cost per head plotted against population density and highlighting the 'good all-rounders' with the red circles
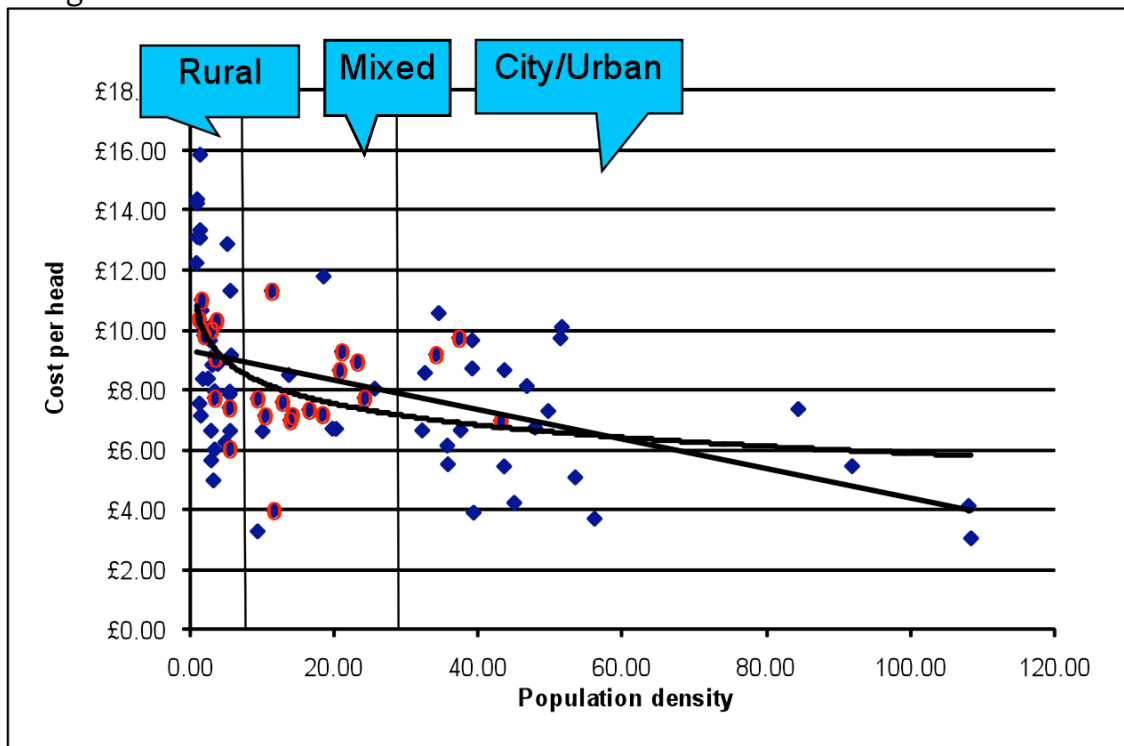
Figure 14 shows the extraordinarily wide range of cases per 1000 of population. Some services have as few as 50 cases per year for each 1000 people whilst other PCTs have a demand that is three or four times as high. What will perhaps be encouraging to commissioners is that twice as many of the 'good all-rounders' are below the line indicating again that the more expensive services are not necessarily the better ones.

Figure 14 Cost per head against the volume of cases per 1000 patients again highlighting the 'good all-rounders' with red circles.
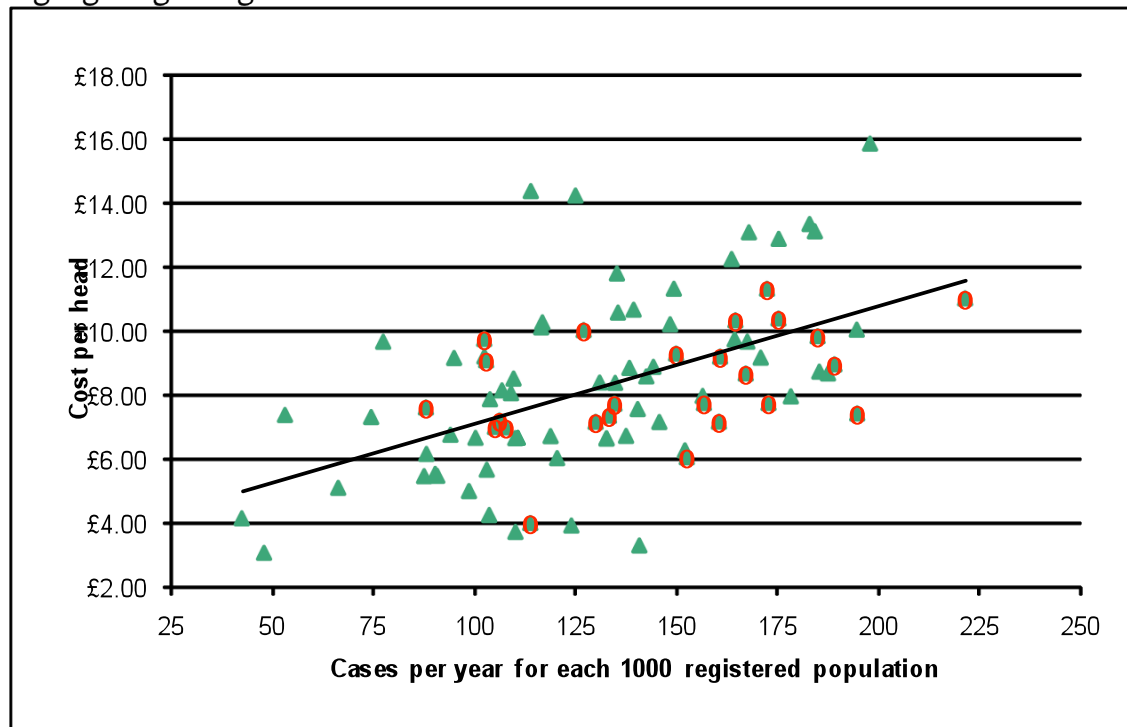
Lessons for commissioners and providers

Commissioners and providers should understand more about demand

We believe that there are some obvious elements in the variation in demand from one service to another. Certain services, particularly in areas to which people choose to retire, receive a higher level of demand because of the higher proportion of older people. Many city areas seem to have a lower demand, which we suspect is because of the plethora of alternatives available to them and perhaps also where some section of the community prefers to use the hospital rather than primary care services. Nevertheless the variation is startling, and we wonder whether in some cases the low apparent demand arises because of a previous poor experience with

the service. Certainly with a mind-set that out of hours care is only for obvious emergencies, and with most contracts effectively for a fixed cost regardless of usage, one could understand how (unconsciously) a service might deter patients from returning.

Other factors do influence cost but the biggest variable is good management

We have included in this report graphs that demonstrate the some of the variability in cost comes from rurality and from the variation in demand. We have, in the feedback events to the two rounds of the benchmark looked at a number of other factors including deprivation, ONS Super-groups, cost of clinical staff etc. Although these must influence the cost (in some cases, such as the cost of clinical staff it is a mathematical certainty), we have seen no obvious relationship in the results.

Having looked at many of these services, we believe that the biggest variable is within the control of the commissioner and provider. Services that ensure that they establish the right number of centres, without having some that are woefully under-used; that ensure that staffing is adequate to meet the predictable demand with a small margin for the unexpected; that ensure that the service is well set up so that staff can work productively; that there is clarity about the duties and expectations for each person operating within the service; that the governance processes support learning; that training is available so that individuals and the service can steadily improve; that ensure staff are not too tired, start when they should and work steadily through their shift; all these seem to us to perform well and offer value for money.

In short it is the detailed management of the service that makes the difference.

Commissioners should expect services to fall in the middle range for cost

When contracts were first let, some services (of all types) under-estimated their costs and the headroom required to invest in developing the service. Some of these issues have been resolved (often because the PCT accepted that the cost would be greater than originally quoted, but sometimes because the service was re-tendered at a higher cost). But there are still a small number of services that are so cheap that they do not have the management headroom to provide a solid governance framework or process. Commissioners might well find that supporting these providers with a little extra resource repays dividends (particularly where otherwise the performance is strong) in allowing the service to develop into a truly excellent service that, when the time comes for re-tendering, will be a strong contender with an existing track record and good local understanding.

Equally we see no reason why services that are towards the top of the cost range should not be challenged, especially when they are not performing as well as others that are cheaper. There will sometimes be good reasons for the higher cost – for example if the specification requires the provider to keep open an excessive number of centres compared with othersor if the specification requires them to provide additional services (though we try to account for this in the benchmark comparison, and ask detailed questions to identify if this is the case). In other cases, however the higher cost may be because the service falls down on some of the details of good management that have been outlined above.

Commissioning

Two years ago, in publishing the World class commissioning competencies the Department of Health made clear that *"commissioners are central to a self-improving NHS"*. Put simply, patients get the service that is commissioned for them.

We agree with the implication from the CQC interim findings that there is more to be done so that out of hours services steadily improve, and that it is the role of the commissioner to drive that improvement. Many of the points have been made above, but it seemed worth highlighting the particular lessons for commissioners and identifying some of the other ways that the benchmark can support them in fostering that "self-improving NHS". Many of the points below will clearly relate to the commissioning competencies, but we will allow the reader to make their own link.

Commissioners should use information and demand detail

There is a host of information that is available about out of hours services including:

- The Ipsos MORI survey provides responses from around 1000 people each quarter about their view of the out of hours service. This seems to us to provide essential information about how organisations are seen by patients as part of any tendering process, to provide a valuable monitoring tool to assess the performance of a provider and can be used to set targets for the improvement by the service. We see no reason why those PCTs where their patients are feeding back information putting the out of hours service in the lowest group in respect of timeliness and in the rating of care received should sit idly by, when other services score almost twice as well on these measures.
- The regular reports from provider services about their performance against the national standards. But commissioners should do much, much more than simply looking at the 'traffic light report' or 'heat map' - they should expect the provider service to prepare and share with them supporting information that allows them to understand the detail such as:
  - Analysis of performance against the main standards by hour of the day to check that the service has matched capacity to demand and that the response is appropriate at all times of the day and night (examples of this analysis for time to definitive assessment were included in the information fed back for validaition).
  - Analysis of priority by individual and of cases where priority is escalated (to ensure that priorities are applied consistently and safely)

- o Analysis of the dispositions, length of call or productivity, referrals to other services by individual (to check that the service is consistent and that the response does not depend on the individual clinician involved)
    - o Increasingly too, services should be looking at analysing common or sentinel clinical conditions, together with a review of how they are treated and how this compares with NICE or other guidelines
- Analysis of the feedback from patients via the patient experience questionnaires. All services are required to collect this sort of feedback, but it is up to the commissioner to check that this is more, much more, than a simple 'satisfaction survey' or 'happy sheet'. In reality it should be a tool that is used to understand the experience of the patient and to gain from them an understanding of how the service can be improved. Questionnaires are, of course, far from the only tool to use for this purpose, but any questionnaire that does not cover the experience of the patient in at least the sort of detail that is used for the CFEP survey which forms part of the analysis of the benchmark will fail to elicit the necessary detail to drive performance improvement. In addition there are a wide range of tools such as focus groups, interviews, mystery shopper that can be used to provide detailed feedback on services. Commissioners and providers might care to look at the video from 'the big brother booth' used by the 5 Boroughs Partnership to understand more about how their services were seen by patients (see http://www.5boroughspartnership.nhs.uk/internal.aspx?PageID=7023)
- Use of the benchmark information. The reports that we produce provide valuable information because they compare services on a like for like basis and give an immense insight into the detail of the service as well. As will be clear from the points made above, we spot many opportunities for services to improve their service to patients, to simplify the operation and reporting and to become more cost effective in delivering a better standard of patient care. As David Colin-Thomé wrote in his letter to all PCTs *"I strongly encourage those of you who have yet to join this benchmarking exercise to do so"*.

No service can assemble this sort of information every month, nor could the commissioners devote the necessary time to it, but they can choose a topic to look at in more detail so that each is covered during the year. It is the commissioner's role to ensure that there is a rounded look at all aspects of the service and this process seems to us to be essential if they are to achieve this.

Commissioners should use this document as a check-list

Within this document we have highlighted a host of variations between services. Some of these are of a technical matter only and it is helpful to understand that the

small difference between the measure of percentage of urgent cases definitively assessed within 20 minutes reported by the service and the benchmark is because one allows for patient attributable delay and the other does not. The difference makes no practical difference to a patient and does not reflect on the quality of the service.

Other differences matter. We have described some services that have a double assessment process or a split service that has an unmeasured delay that is not reported upon and that affects the service that is received by all or most patients. We have described services that appear to be unable to report even against the national quality requirements and that have systems or coding that make it impossible to achieve that deeper analysis that is so essential to ensure that the service is both safe and able to improve its performance. We have also described services where we have real questions about whether the service has properly ensured that an urgent case will be recognised by the non-clinical call-handler and where we suspect that the service should be judged against the requirement to assess all cases within 20 minutes.

We have also highlighted in each of the reports, and again in this document, how the list of questions about governance included as an appendix to the report provides a really useful way for commissioners to understand the governance processes that are in place to improve the consistency, reliability and quality of the response by the out of hours service.

Good commissioners will already know enough about their provider service that it should not take long to use this document as a check-list to ensure that the concerns that we raise do not apply to their service. Indeed several of these commissioners and providers have already checked any anomalous result or comment in our report on their service to be sure that they understand if action is necessary by them or the provider. Commissioners that are less good may find that it takes them longer, but we believe it will be worth it to assure themselves and the patients of a good quality service that is able to be part of the "self-improving NHS".

Commissioners should not be afraid to act

What does concern us is that some commissioners seem to be apathetic or happy to wait for the contract to be retendered in the hope that things will be changed by re-letting the contract. In other cases the out of hours service reported to us that there was no consistency in who took responsibility for the commissioner in managing the out of hours arrangements - indeed commissioners themselves have described how project managers were changed even part way through the tendering process.

We are firmly of the view that it is the responsibility of a commissioning PCT to work with a poorly performing service and to insist on action being taken immediately. In every case that we have described of poor performance (failures of systems, failures in reporting, failures in integration with other services) there has been more than enough evidence to allow the commissioner to terminate the contract. We don't think that they should, but they do have more than enough levers to demand that an immediate improvement plan is put into place and to insist, if they conclude that the provider management team is incapable of fixing the service quickly (whether the provider is independent or part of the PCT), that suitably qualified interim managers are found who will address the shortfall.

Although they are far from being the majority, it seems to us that if even a few commissioners fail to engage adequately with their out of hours provider service, fail to make use of the information available to them, fail to demand detail beneath the headline information and fail to spot the indications of poor service or performance from our benchmark report, those few are far too many. We suspect that there may be more than a few. We would be delighted to be contacted by more who wanted to discuss the detail of the findings and understand our perception of the opportunity for improvement.

Future development of the benchmark

The first three rounds

We have never seen the benchmark as something that is static. In the first round we established the process and many of the basic measures, highlighting some of the differences in reporting that we have described earlier and checking the codes to ensure that we were measuring each service accurately.

In the second round of the benchmark we looked particularly to include the voice of the patient drawing on the findings and analysis of the CFEP survey that we commissioned and the findings of the Ipsos MORI GP survey, comparing patient perceptions with the measured performance of the service.

The third round will be focused narrowly on performance over the Christmas period and will look in more detail at some aspects such as the escalation of priority and will look at productivity in a different way. We feel that it will be valuable to look at this period when all out of hours services are at their most stretched, particularly because it will throw light on the operation of the overall urgent care system.

Developing the comparison further

We are also looking to develop the benchmark in a number of ways before the fourth round planned for mid 2010 which will again include the patient experience questionnaire. We will be looking particularly at:

- Enhancing the benchmark to cover the wider case mix of integrated services. It may take a little time before providers have tightened up their coding so that they can reliably separate cases for this purpose and for PCTs to be able to realistically cost the different types of service provided separately. Nevertheless this will be valuable and will allow services that are integrated in a variety of different ways - some with walk in centres, some with minor injury units, some to accept emergency department patients, some with GP Health centres - to compare performance, effectiveness and cost-benefit not just of the out of hours service but of the other elements as well.
- Modifying the questions used about some of the governance processes - looking particularly at the recommendations from the CQC findings in relation to the Cambridgeshire TCN case.

- Developing ways to include a comparison between services about any planned clinical working hours that were unfilled and looking to develop the governance questions (that we believe provide such a good check-list that services can use to develop their own governance processes) to include questions about how checks are made on compliance with the working time directive.
- Developing a process whereby services can look more closely at the detail of how drugs are managed and prescribed and compare good practice. We have already agreed to work with the NW medicines management network to do this.
- Involving PCTs in the review of the service, particularly by using them to moderate the answers to the governance questions so that they seek evidence and ask for examples to understand the feedback to clinical and non-clinical staff. We are certain that being involved in this process will not just improve the comparability across services in the benchmark but will ensure that PCTs become much more involved in looking *at details like the quality of clinical decisions, the efficiency of call handling, the adequacy of staffing*"
- Organising an additional feedback event targeted at analysts and those involved in reporting. We have already flagged this at the five events following the second round,but we think that it will probably support providers in understanding how they can address many of the points made in this report.

Opening the benchmark up

Finally we are planning to open up the information to allow PCTs and providers to identify other services. We are certain that this will improve the benchmark and drive up performance further.

Understandably, providers and commissioners have taken time to gain confidence that the comparisons that are made through the benchmark process are reasonable and that the explanatory factors and caveats are reasonably described.

With help from participants to make sure that we get it right, and with participants understanding the importance of validating the analysis carefully, there has been strong support for moving to openness over the main comparators for the full benchmark in the summer of 2010 which, importantly, will include the patient experience information that provides such an important counter-part to the analysis of cases, performance, cost etc.

There will be benefits in that poor performers will have fewer places to
hide, good performers can demonstrate their effectiveness and we are certain that,
being able to identify comparable performers will convince many services that they
can do things better - and once convinced of this, they will.

All of this will involve additional work so that providers and commissioners can
manipulate the information about each service, but we are determined to continue
to develop the benchmark to support the thriving and fast developing out of hours
services to do even more for their patients.

Conclusion

We set ourselves an ambitious target for this report to encapsulate all of the learning from the detailed analysis from the first two rounds of the benchmark. The report is aimed at both commissioners and providers. Commissioners, because it is they who have the main responsibility for ensuring that the service provided to their patients is safe, responsive and supportive of patient needs whilst also offering good value for the tax-payer's money. Providers because much of the detail will need to be understood by them if they are to be able to respond and improve the service where necessary.

Ideally we would have liked it to be shorter, but we felt that to leave out the explanatory detail and not to provide at least some guidance on how things could be done better would negate its value.

Despite this length and our best endeavours, there may be elements that are not clear, or services may wonder if specific aspects are relevant to them. Our aim is to help individual services and the sector as a whole (which is already ahead of others in the way that it is managed and the service that is provided to patients) to become an examplar of what can be done.

We would be happy to support services in understanding more, initially with a conference call that should involve both the commissioner and provider. To be of most value we would want to prepare for the call and would ask commissioners to contact us in advance by email, describing any specific aspects that they wished to discuss, and providing some of the recent reports from the provider service that have been used as part of the routine monitoring of the service together with further details.

To arrange this, or to comment on any aspect of the report, please contact Henry Clay, preferably by email (henry.clay@primarycarefoundation.co.uk) but otherwise by phone on 07775 696360.

Appendices

Coding – categorisation of cases

:

Case Type

Use case type to identify the main category of outcome into advice, base or home visit. By all means sub-divide it into nurse advice, doctor advice or in other ways that are meaningful to you. However we would recommend that you avoid using this to identify referrals from the MIU, WIC, A&E etc (see below for alternative suggestions), for identifying different services such as the WIC or treatment centre that you may run. A particular difficulty that many services have is that they set up a case type of District nurse which is then used for two completely different purposes. The call-handlers use this code to indicate that they handled the call and passed it onto the district nursing service, whilst the clinicians use the same code to indicate that they referred an out of hours call to the district nurses. Because of these different uses it becomes impossible to distinguish the calls that came in for the district nursing service from those that came in as OOH calls.

Call origin

The coding of call origin often becomes horrifically complicated. I cannot believe that it is easy for users to have to choose from 40 or more codes,and I cannot believe that the coding accuracy is adequate to allow any meaningful analysis. There are two aspects to this coding: one is about the relationship of the caller to the patient, and thisincludes mother, self, parent, daughter, etc. If you want to analyse your calls in this way, then you need to think what small number of codes are useful so that the clinician knows who they are speaking to and for analysis purposes – but I would have thought that a maximum of 8 or so made life easier for the users. The second aspect is when the list includes calls from the ambulance service, and has then been extended to identify different types of service by using it to distinguish walk-in cases, referrals from A& E, referrals from the WIC or MIU, calls from patients wishing to contact the district nursing service, calls from the hospital laboratory about test results that need following up. Mixing these two different categories causes confusion. Typically some people will record a patient that arrives as a walk-in or A&E referral as 'self' and others will use the 'walk-in' or 'A&E referral' code. The message here, as elsewhere, is that you should decide what you want the code to be used for and use it ONLY for that – making sure that the list is small enough that it is easy for users and that they understand when to use which code.

Service

The Service field available in the more recent releases of Adastra is potentially very useful. Hardly any services are using it yet. Not only can you use it to separate (for example) your WIC cases from your OOH cases from your A&E referral, but you can also set up the workflow and choices differently for each of the different services. A simple example will illustrate the point. If you ran an OOH service and also operated from a hospital providing an Urgent Care Centre closely integrated with A&E, you might choose to set up two services, one for OOH and one for the UCC/A&E operation. Whilst you might use the OOH case types to identify Advice, Base visit and Home visit, and the call origin to identify the relationship to the caller (parent, self etc), you might have a different set of case types for the UCC to identify assessment, X-Ray, Blood tests and use the call origin to identify whether the patient walked into the UCC or was referred from A&E. Because the Service field effectively allows you to separate the two services, users in one area will not see the choices available to others except where you want them to – thereby simplifying the task for the user. I am not an expert in what can or cannot be done using this field,but Adastra can certainly advise you on how to exploit this functionality to simplify the operation and reporting when you run several separate services. This function may also be useful even if you only have one type of service if you share a hub, because it may allow you to separate out your own codes from those used by other provider services.

Priority

Most providers use the coding for priority in a fairly simple way identifying emergency, urgent and less urgent cases in a number of slightly different ways. However there are a small number of services that have abused this code to identify MIU/WIC/A&E/Surgery calls or to identify types of service such as District nursing or non-conveyed ambulance calls

Doctors operating group, Health board and commissioning group

These are described together because they all allow different groupings of practices for different purposes. these codes are all set up to link to the patient's GP practice . A logical way of using them might be to identify the PCT using the doctors operating group, and to distinguish whether the practice has opted out or opted in using the commissioning group,leaving the health board for any future use. Some services have used this category to separately identify (for example) Dental calls from OOH Cases from District nursing calls or to separately identify telephone answering services. What they have had to do to allow this is to establish a practice number (or use the PCT number) and then code each case accordingly. This means that the important information about coding a case to a patient and to their practice is lost. Importantly you do need to identify which practices are part of which PCT, if you are

to be able to report by PCT. This weakness was highlighted by the
Healthcare commission report,but what they did not see and we have seen in the
data extracts is how many practices are sometimes misallocated to the wrong PCT or
how practices can appear to be in two PCTs, presumably because the fact that a
doctor has moved practice has not been reflected on the system

Informational outcomes

Informational outcomes provide a good way of classifying the outcome for analysis
purposes. We have suggested in the past that these should include ALL of the
following (sub-divided if essential).

- Ambulance/999

- No further action

- Patient to attend own surgery

- Referred to A&E or MIU

- Referred for admission

- Referred to other services

- Recommended for follow-up by GP

- Self-referred to Hospital and, possibly

- Patient Deceased

We recognise, however that in certain areas there may be different working
arrangements for referral and the PCT may want to clearly differentiate referrals to
primary care from secondary care so an alternative might be to choose the following:

- Ambulance/999

- Chose to attend hospital

- Episode of care ended – no further action expected (by any organisation)

- Own doctor – patient to book appointment

- Own doctor – recommended for follow-up please

- Primary Care – referred to WIC/MIU

- Primary Care – referred to other service (rapid response etc)

> ➢ Secondary Care – referred to Emergency Department

> ➢ Secondary Care – referred to other department

> ➢ Social services – referred for support

> ➢ Tertiary Care – list local examples

We strongly advise against using informational outcomes to classify the case for other purposes. Users have sometimes used this as a form of clinical grouping (fall, asthma, breathing difficulties, diabetes etc.) or have used possibly meaningless or dangerous phrases such as 'satisfied with treatment' or 'safety measures advised' (in the latter case there is a danger that the lack of such an informational outcome might be taken as implying that the clinician did not provide the appropriate safety advice).

Calculation of performance against telephony standards

These worked examples are provided to describe how we calculate the performance against the national quality requirements for telephony (QR8). The service needs to know the number of engaged calls and needs to discount calls abandoned within 30 seconds (or within 30 seconds of the end of the message) which do not count in the calculation. However the service should count all calls that are abandoned after this time and should calculate the percentage of calls answered within 30 seconds (or 60 seconds of the end of the message) using <u>all</u> calls including those that have been counted as abandoned.

The worked examples are deliberately simplified, though in most cases the telephone system will hold an effectively unlimited number of callers waiting for a response so that there are no engaged calls.

Example 1 with no message

1085 calls hit the switch, 5 get an engaged tone and 80 abandon within 30 seconds, but another 20 abandon after that point.

5/1085 = 0.46% engaged. None of the first lot of abandoned calls are counted as abandoned or counted in the divisor so the calculation is 20/(1085-5-80)=2.00% of calls abandoned.

Of the 1000 calls that the service should have answered (because they did not get an engaged tone and were not abandoned within 30 seconds) the service answered 923 within 30 seconds so the performance is92.30% answered in so seconds

Example 2 with a message that is 20 seconds long

1105 calls hit the switch, 5 get an engaged tone and 100 abandon either during the message or within the next 50 seconds. But another 20 abandon after more than 50 seconds (including the message time).

5/1105 = 0.45% engaged. None of the first lot of abandoned calls are counted at all so the calculation of abandoned calls is 20/(1105-5-100) = 2.00% of calls abandoned.

Of the 1000 calls that we should count (because they did not get an engaged tone and were not abandoned during or within 30 seconds of the end of the message) the service answered 943 within 60 seconds of the end of the message so the performance is 94.3% answered in 60 seconds of the end of the message.