FINAL REPORT

June 2016

An Independent Review of Emergency Road Ambulance Service Funding

Prepared by Dr Murray Horn

Commissioned by

the Director-General of Health and the Chief Executive of ACC

Contents

Executive summary

The Ministry of Health (the Ministry) and the Accident Compensation Corporation (ACC) jointly fund about 70% of ambulance services provided by the Order of St John (OSJ) and Wellington Free Ambulance (WFA). This funding is based on the actual cost of ambulance services in 2012 increased by a general cost index every year.

This arrangement is not sustainable because:

a. both providers cannot continue to absorb the cost of real demand growth and so will have to keep coming back for extra funding (OSJ's demand growth has been about 4% per annum for the last two years, while WFA's demand has been reasonably flat[1]); and

b. the distribution of the funding does not give enough recognition to the higher costs that OSJ faces in providing ambulance services to its larger rural population. OSJ has been able to manage on a substantially similar age-adjusted per capita funding as WFA by relying on a combination of extra revenue efforts (e.g., user charges) and cost savings (e.g., single crewing, largely in rural areas). OSJ does not believe it should continue to rely as heavily on single crewing.

The two providers are also looking to expand the range and quality of services that they offer (and securing some funding for these additional services) and to recover more of the cost of providing existing services. The Ministry, and ACC to some degree, can help facilitate both objectives.

This report suggests five main changes to funding and accountability arrangements that will help address the two issues listed above and, therefore, help meet the Ministry and ACC's desire for government funding arrangements that will provide quality, innovative and sustainable services.

1. Setting conditions for funding: The providers know that the funders have to meet the cost of any provider decisions or omissions that threaten the viability of the ambulance service, at least up to a point. This moral hazard can be reduced by making it clear that funding will become increasingly conditional on the provider taking specific actions to improve their financial position, without recourse to the funders, as their ability to manage financial risk deteriorates (e.g., as reserves are run down or on-going deficits are forecast).

2. Funding demand growth: This conditionality will not work if it is inevitable (e.g., because annual funding increases fall too far behind annual demand growth). Funding should be automatically adjusted to meet real emergency service demand growth that cannot be reasonably met through provider efficiency gains. While the providers have little control over emergency service demand, they can influence the cost of meeting that demand.

[1] National Ambulance Sector Office's (NASO's) actual (January 2013 to March 2016) and forecast (April to June 2016) incident data indicates average national demand growth of 3.4% per annum from 2013/14 to 2015/16, with OSJ at about 4% and WFA basically flat. A Sapere Research Group report (2015) projected OSJ demand for of about 4% over the following five years, with growth of about 2.6% over the previous five years. Research by the management consultancy research company ORH (2014) expected WFA demand to grow by about 1.4%, although it noted that demand had been relatively flat for WFA since 2009. OSJ reports 111 call growth of about 4.6% for the 2014/15 year and projects call growth of 6.2% for the 2015/16 year, with incident volumes lower (circa 4%).

The funding formula needs to help ensure better value for money by encouraging both the right timing of investment in extra capacity and the right mix of capacity. Encouraging the right investment timing requires a different approach to funding rural and urban capacity. Encouraging the right capacity mix requires the funding to encourage the providers to provide the right type of response to the right patient at the right time. While the resulting approach involves a number of elements, these can be summarised in a formulaic approach to funding rural and urban services that will provide much greater certainty to both the funders and providers over time.

3. Funding a rural loading: A rural loading should be added to the current funding base, primarily aimed at significantly reducing the practice of single crewing in rural areas. While the exact size of this premium is a matter of judgement, there are a number of reasons why the amount for the first three years should be less than the level that would allow for full double crewing. Additional funding of about $10.6m should be applied over the first three years to add capacity where it will have the greatest impact in both meeting likely rural demand and reducing the number of single crewed responses in rural areas.

4. Increasing funding from other sources: Two areas of change can help increase revenue from other sources.

   First, increasing the scope for third-party funding of emergency road ambulance services (ERAS) by: (a) reducing the cross-subsidy from ERAS to non-emergency ambulance services by requiring organisations that routinely make non-emergency calls to meet the full cost of the services they use and (b) increasing the maximum user charge permitted for individuals that use ERAS.

   Second, there are a number of additional value-added services that ambulance providers can supply to ACC (injury prevention), primary health organisations (PHOs) (after hours rural services) and district health boards (DHBs) (reduced unnecessary hospitalisations; improved survival and return to function for a few key conditions, such as trauma and heart attack). The Ministry and ACC can help ensure that services that pass the usual business-case tests are funded, either directly or via the beneficiary PHO or DHB.

5. Changing funding of communication centres: Finally, current funding for communication centres could be incorporated into each provider's funding base rather than funded separately after a fixed contribution from each provider. While this does not change the initial quantum of funding provided by ACC and the Ministry, the quantum would increase on the basis of the suggested annual adjustments. The providers would then face the full marginal cost and benefit of the operational and investment decisions they take (including the potential for communication centre amalgamation).

These five changes will place ambulance funding on a firmer foundation and should reduce the providers' forecast operating deficits. While this should also help the providers decide how to tackle existing deficits, some funding conditions may well be required for OSJ while the new arrangements are introduced, given the size of its deficit (i.e., conditions aimed at improving OSJ's financial position without recourse to the funders; from operational efficiencies through to asset realisation).

1. Purpose

This report is the result of a request from the Director-General of Health and Chief Executive of the Accident Compensation Corporation (ACC) to provide them with advice on government funding arrangements that will provide quality, innovative, and sustainable emergency road ambulance services (the Terms of Reference are attached as Appendix 1).

2. Current funding environment

The main features of the current funding environment are listed below.

a. Emergency road ambulance services (ERAS) are provided by the Order of St John (OSJ) and Wellington Free Ambulance (WFA). WFA provides services for the areas covered by Capital & Coast, Hutt Valley and Wairarapa district health boards (DHBs), while OSJ covers the rest of the country.

b. The Ministry of Health (the Ministry) and ACC together cover about 70% of total ambulance costs with the remainder of funding coming from DHBs, donations, commercial profits, volunteers and, in the case of OSJ, part charges. ERAS costs are only part of the total cost of ambulance services (e.g., patient transfer services share common infrastructure).

c. The Ministry and ACC now both fund ERAS capacity through joint service agreements, whereas ACC used to fund on the basis of transports to hospital. The National Ambulance Sector Office (NASO), a business unit within the Ministry, manages the funding arrangements and service agreements for emergency ambulance services on behalf of both ACC and the Ministry.

d. Current service agreements require the Ministry and ACC to pay a sum based on the actual cost of ambulance services in 2012 increased by a general cost index every year. Demand for emergency services via the 111 system is growing at about 4% per annum for OSJ, with little growth for WFA. The two providers are expected to absorb the cost implications of real emergency service demand growth and improve or maintain their performance against a range of measures, year-on-year until 30 June 2017.

e. OSJ and WFA are both transforming their business models: from being largely a response-and-transport-to-hospital service to becoming more capable and qualified mobile providers of health services, e.g., by providing more services over the phone (hear and treat); by paramedics providing health services at the patient's location or in partnership with primary and community providers (see and treat) and by transporting patients to primary or community care destinations.

The two providers argue that this funding arrangement does not provide them with financial sustainability and certainty for the future given on-going demand pressure.

3. Objectives of the report

The main objective of this report is to suggest how Ministry and ACC funding and associated accountability mechanisms might be reconfigured to provide greater financial certainty for both the funders and the providers while encouraging the provider business models to continue to evolve to deliver more value for money.

A secondary objective is to consider the scope for both recovering more cost and generating more revenue from existing capacity by delivering additional value that will help ACC, the Ministry and DHBs deliver on their wider objectives.

4. Overview of the report

The rest of this report is presented in six sections.

Section 5 describes the underlying funder-provider relationship, which determines the choice of funding and accountability arrangements. In this case, while the arrangements comprise many elements of an arm's-length contract, there is a reasonably high degree of co-dependency between the funders and the providers that needs to be taken into account.

Section 6 is the main body of the report and develops the recommended approach to funding ERAS capacity that puts WFA and OSJ on a comparable basis, shares both the inflation and the real demand risks between the funders and the providers and better aligns funding to support the desired shift in business models.

Section 7 provides a quick description of the recommended approach to funding value-added services that would fall outside the capacity funding model described in section 6.

Section 8 responds to the requirement in the Terms of Reference (see Appendix 1) to reconsider funding of the communication centres, describes the options and recommends a way forward.

Section 9 summarises OSJ and WFA feedback on the initial draft recommendations that follow from the approach suggested in this report.

Section 10 summarises the main arguments and final recommendations, taking provider feedback into account.

5. The funder-provider relationship and accountability

The nature of the relationship between the two funders (the Ministry and ACC) on one hand and the two providers (OSJ and WFA) on the other is fundamental in determining the choice of funding model.

5.1. Provider autonomy is desirable and requires financial flexibility

The providers want to remain autonomous organisations with the management discretion that implies. This allows them to manage their organisations and serve their communities in the way they think best. Such autonomy helps to maintain a community-based brand that attracts volunteers, sponsorship and community funding and supports the provider's commercial operations. It also helps build community resilience. A strong provider brand is also in the interests of the two funders, as long as they are not being asked to subsidise other activities to support this branding.

The quid pro quo of this autonomy is that OSJ and WFA are able to live within their means, i.e., that they do not come back to the Ministry and ACC to fund poor decisions or cover financial risks that have not been well managed. They must maintain enough financial capacity to absorb these risks themselves, e.g., by drawing on other sources of income, by managing down their costs, by calling on their reserves and by managing their assets (including via asset realisation, if need be).

This allows for a reasonably arm's-length funding relationship, where the funders commit to a funding track that will support the providers to deliver an emergency response service to their populations. The providers are then free to manage their organisations in a way that delivers the agreed levels of service for that capacity funding, as well as anything else they may want to do.

5.2. Co-dependence creates moral hazard which needs to be addressed upfront

While this arm's-length model is at the core of the funder-provider relationship, the reality is that there is a greater degree of co-dependence between the funders and the providers than the model implies. The providers cannot exist without Ministry and ACC funding, and it would be both disruptive and expensive to the funders if one of the providers failed.

While OSJ may well be able to assume management or even ownership of WFA if the latter failed, the opposite is unlikely. The most likely alternative to the failure of OSJ (and possibly WFA) would be for a publicly-owned entity, like a DHB, to take over ownership of the service. That is because of the need to maintain service continuity and because the circumstances that lead to provider bankruptcy would likely discourage other private investors from stepping in quickly. While the services agreement does recognise the risk of provider bankruptcy, changing provider ownership would impose real costs on the funders; thus the funders have an interest in maintaining provider solvency.

In short, up to the point where change of ownership is the best option, the existing providers are, in a sense, too important to fail, with all the moral hazard this implies. Unless this co-dependency is addressed upfront, there is a risk that OSJ and WFA could assume that decisions or omissions that put the ERAS at risk will be funded as long as they are not so bad that the funders decide that the services would be better being owned by someone else.

5.3. Accountability arrangements need to address co-dependence

The way to address this co-dependency is to manage the basic arm's-length funding arrangement inside a strategic relationship that is based on a combination of full provider disclosure and funding conditions. In this instance, 'funding conditions' means that funding becomes conditional on the

provider taking specific actions to improve its financial position, without recourse to the funders, as its ability to manage financial risk deteriorates.

5.3.1. Full disclosure

Disclosure should not become a burden for the providers, and it is unlikely that NASO would require regular disclosures that were more demanding than those required by the providers' own Boards; the providers' Board papers could simply be copied to NASO.

Disclosure should at least include key operational, service and financial performance metrics as well as the normal quarterly financial reporting and draft forward financial and operational plans (so they can be discussed with NASO before being set in stone).

5.3.2. Funding conditions

Imposing funding conditions implies increasingly more direction from the funders as the provider's financial flexibility is eroded. In order to create the right behaviours, funding conditions should be avoidable by good operational and financial management by the providers.

However, they will be unavoidable if the funders expect the providers to absorb all the costs associated with real demand growth. Thus the funders need to share some of the financial burden of real demand growth. While the funders should expect the providers to manage some of this demand growth by improving their efficiencies, it is unrealistic to expect on-going efficiency gains of 4% per annum. The funders have to share some of this burden – and do so in a reasonably predictable way – before it is justifiable to impose any funding conditions.

Funding conditions require:

a. a clear set of parameters that will trigger each stage of the conditions: these triggers need to be reasonably well understood by the providers so that they know how close they are to triggering the imposition of conditions

b. a clear set of actions that the funders require of the providers when the triggers are breached: actions that will improve the provider's financial position without recourse to the funders.

The right set of parameters for point a. above will be a combination of operational, service and financial metrics. While these should define the degree of financial flexibility that the providers have to absorb financial risk (as a bank might require before extending credit), operational and service metrics need to be monitored to ensure that the flexibility is not being created at the expense of service quality or sustainability.

Forecasts of on-going operating deficits and/or a reduction in liquid assets (reserves) below levels required for normal operating purposes would be the sort of reduction in financial flexibility sufficient to trigger the imposition of funding conditions. Something akin to a credit rating process would provide a good signal of the current level of comfort or concern.

In terms of the conditions themselves, a simple two-stage process would probably be sufficient.

- Stage 1: This stage would be triggered by a reduction in financial flexibility that should be able to be corrected if the provider makes operational or financial management changes that they have not made and would be required to make by the funders.

- Stage 2: If the actions in stage 1 were insufficient to restore financial flexibility, then stage 2 would require the providers to seek and obtain funder approval for their budget until their financial flexibility was restored. If budgetary approval was insufficient to restore financial flexibility then, at some point, the funders would need to have a conversation with the relevant Board about preparing for a change in management and/or ownership of the organisation.

The first stage may involve a combination of actions that are sequenced to restore the desired degree of flexibility and would include some, or all, of the following options.

- Improving operational efficiency: This requires a credible process for operational benchmarking, and for changing models of care, that can identify where the most valuable efficiencies are likely to be found.

  For example, the work that management consultancy research company ORH[2] completed for WFA indicated the scope for significant gains (e.g., in reducing activation times) that could be translated into new investments for capacity in particular geographies that together allowed WFA to meet increased demand with minimum increase in resources and cost. The providers would have to agree to a plan with the funders to realise these efficiencies over time (in the form of a lower cost to meet future demand).

- Merging or rationalising functions to reduce cost: The two most obvious examples would relate to the providers sharing back office services and/or rationalising communication centres. The discussion of communication centre funding in section 8 of this report recommends that these be no longer separately funded, requiring the providers to weigh up the savings from merging this function against the benefits of retaining greater control. However, this assumes that providers can afford the luxury of forgoing what seem significant savings. If the cost of retaining control effectively increases the financial risk to the funders then, once the triggers for this stage are met, on-going funding could be conditional on call centre rationalisation.

- Reducing cross-subsidies and improving user incentives to help reduce/meet the cost of increased demand: Where appropriate, the providers would need to look for ways to introduce or increase existing charges (especially for commercial users of the service), identify and reduce cross-subsidies to other parts of the business and manage frequent users and responses for people with chronic conditions more proactively (e.g., chronic obstructive pulmonary disease (COPD), cancer, asthma, diabetes and palliative care).

  About one-quarter of all calls to communications centres are through the providers' non-urgent lines (i.e., phone lines other than the 111 and staff lines). The funders would want to know more about who is making these calls, what their incentives are and what the acuity profile of these calls look like to ensure they are not being cross-subsidised.

  Anecdotally, a lot of calls from commercial users are of low priority (e.g., OSJ estimates that other health providers – including GPs and aged residential care (ARC) facilities – generate nearly 25% of their ambulance demand). To some extent, an ambulance service that is free to commercial users will substitute for alternatives that impose a cost on those users (e.g.,

[2] ORH. April 2014. Wellington Free Ambulance: Review of the emergency and patient transport service.

having more staff on premises after hours in ARC facilities). It is, therefore, much more important that these users face the full cost of using ambulance services than it is for individuals and families who are not faced with the same range of commercial options.

Requiring commercial users to meet the cost of providing ambulance services may require Ministry support to encourage them to use ambulance services only when such services are really required rather than just because they are free.

DHBs should be required to meet the full cost of ambulance patient transfer services (PTS), rather than shift some of this cost onto the ERAS (e.g., OSJ estimates that the operational deficit on its provision of PTS is about $1.7 million per annum). The Ministry could play a useful role in facilitating this outcome. For example, OSJ suggests that the Ministry manage a national contract for the PTS service that set PTS prices and individual DHBs purchase the service on the basis of these nationally determined contract terms. Again, making users face the true cost will encourage better decisions about ambulance use, including the timing of that use.

The current arrangements limit the part charge that the providers may levy on service users. Users can avoid this charge by paying an annual amount, and many do. The providers have to balance the extent to which increased costs should be covered by part charges against any negative impact on their brand value (and thus on community donations and volunteering) as well as on the non-collection rate. This provides a real constraint on the providers' willingness to increase these charges (indeed, WFA does not impose any charge and, as its name suggests, makes this a key element of its brand value).

Setting the right part charge is a decision that the providers are, therefore, well placed to make. They should be given much wider scope to set these charges.

The funders also need to be reassured that non-ERAS currently provided free or at highly subsidised rates to communities are necessary to maintain brand value.

- Capturing existing brand value: The providers earn revenues that flow from the value of their brand, and a large part of that brand value derives from the supply of ERAS. Some proportion of the revenue that the providers earn should, therefore, be used to fund ERAS.
- Asset and reserves management: Reserves and other assets provide financial flexibility that may need to be used to cover temporary shortfalls in revenue from other sources. Once trigger points are reached, the funders may want an influence over the provider's asset and reserve management decisions.

If the result of imposing a combination of some or all of these conditions proved insufficient in restoring financial flexibility to a reasonable level, then the next stage of conditions for funding would be triggered, i.e., the requirement for the funders to approve budgets.

6. ERAS capacity funding track: certainty and value

The recent shift from funding activity to funding capacity makes sense. However, this approach needs to be developed to provide more funding certainty at the same time as aligning funding to better support the provider's desire to invest in the right sort of capacity at the right time.

Current funding arrangements are based on the real cost of maintaining existing capacity (i.e., adjusting for a measure of economy-wide inflation). That requires renegotiating funding arrangements regularly as capacity needs to increase to meet increasing demand at existing service levels. Having to come back to the funders for more money creates unhelpful funding uncertainty for both the provider and the funder.

Future funding would be more predictable if there were a more regularised and consistent approach to funding increases in capacity. Greater certainty gives the providers more confidence to invest as well as reinforcing the need for them to manage the risks they are best placed to manage, rather than assuming they can come back for extra funding when those risks arise.

Funding arrangements could also be better aligned to encourage more value from the available capacity. Both the providers are looking to better match the services they provide to the needs of those who call for assistance, i.e., providing the right care to the right person at the right time and place. The providers recognise that not everybody needs to be transported to an emergency department (ED). Indeed, this is one of the reasons for ACC's shift from funding transports to ED to capacity funding. Some patient need can be better met over the phone (hear and treat), by paramedics on the spot (see and treat) or by transporting to primary or community care rather than to the hospital ED.

6.1. Funding capacity as insurance cover

Funding to maintain an ERAS capacity to respond to emergencies is analogous to providing the population with emergency response insurance cover. Service coverage is defined largely in terms of response times for urban and rural populations that vary by the priority accorded the emergency.

The Ministry and ACC funding is effectively a premium paid to maintain coverage of a certain quality on behalf of the population being served by each provider (with a maximum individual user co-payment). Treating capacity funding as a premium provides a useful basis for varying funding in a systematic way over time; one that can be applied equally to both providers.

Any increase in the premium that the Ministry and ACC pay should be based on changes in the volume of emergency demand[3], ideally weighted by incident priority. While many factors will influence demand, the main drivers are population size and age (approximated by the proportion of the population over 65 years of age). Population age is important because ambulance use in the over-65-years group is higher than it is for the younger population, and this group is growing more rapidly. Allocating base funding between the two providers based on demand growth will reflect changes in the characteristics of the populations they serve.

[3] See section 6.4.2: Non-rural (urban) services for details on how to measure demand based on conclusions about communication centre funding arrangements (which influence the importance to be placed on calls as well as incidents).

Weighting incident volume by priority mix is useful for two reasons. There is likely to be more value in responding to a life threatening and time-critical incident (assigned a purple or red priority) than to a less acute and time-critical incident (orange, green or grey priority). More explicitly, recognising this value in the funding mechanism also better aligns funding to the providers' desire to better match their response to need (e.g., the scope for both 'hear/see and treat' responses is far broader in the less acute and less time-critical categories).

While funding on the basis of population driven demand makes sense longer term, the way funding adjusts to this long-term position should vary to encourage efficient capacity investment.

6.2. Using funding capacity to encourage efficient investment

The timing of capacity investment is one of the most important decisions that a provider will make: investing too early increases cost unnecessarily while investing too late risks degrading response times.

In order to have the capacity to respond to unforeseen events within agreed response times, the providers expect a significant amount of capacity to be unutilised for long periods (e.g., OSJ plans its capacity investment to target 55% capacity utilisation rate, although the actual utilisation rate will vary with demand).

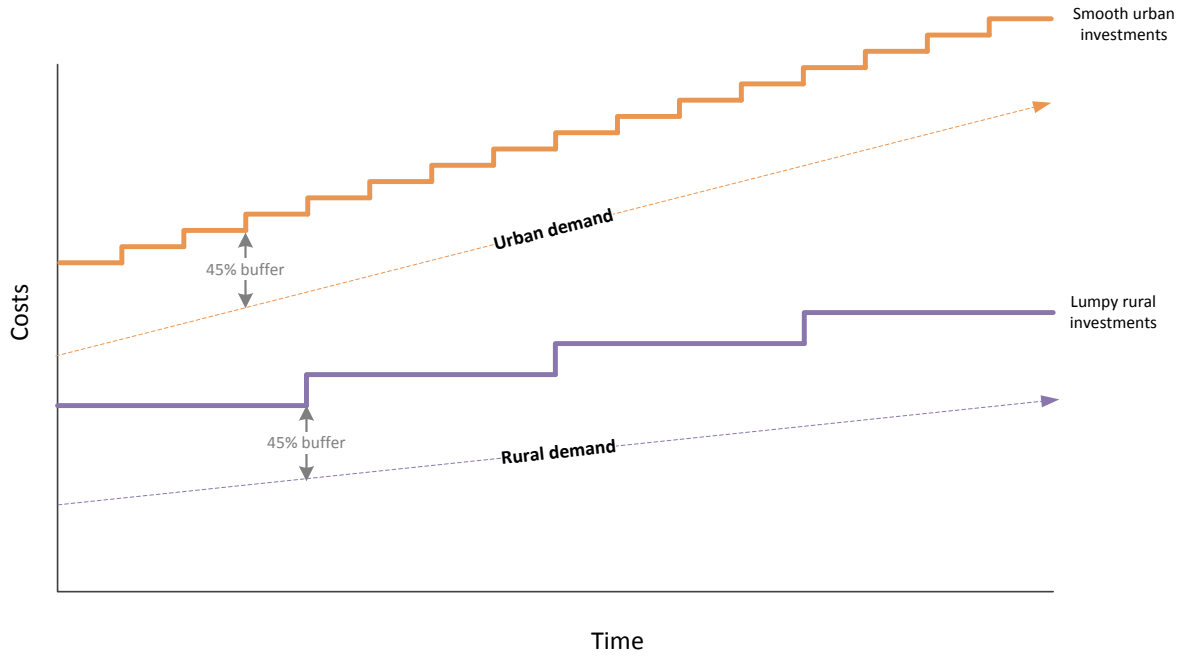
Over the longer term, it should be possible to vary capacity to match demand and so reach the desired capacity utilisation rate and associated response times. Indeed, if capacity could be adjusted quickly to meet demand, then tying funding to demand would also encourage the right timing of capacity investment decisions in the short term.

However, the ability to match capacity to demand in the short to medium term will vary. Capacity has to be added in discrete 'lumps', albeit to varying degrees of 'lumpiness', as existing capacity utilisation rises to the target level (i.e., from extra personnel, extra shifts, extra ambulances to, at the extreme, extra stand-alone stations).

The more densely populated an area, the greater the ability to spread each of these lumpy investments over a larger population base. So, for densely populated areas, it should be possible to provide fine variations in capacity to align with changes in demand and therefore stay close to the planned utilisation rate and associated response times at any point in time. In these cases, funding capacity growth will be well approximated by funding demand growth.

However, for many areas it will be impossible to match capacity to demand in the short- and possibly even medium-term. This problem becomes more acute the more sparsely populated the area. Lumpy capacity investments will either lead or lag changes in demand, depending on how growth is funded and the flexibility the funders extend to the providers in meeting agreed response time goals. This difference is illustrated figuratively in Figure 1.

Figure 1: Lumpy and smooth investment to manage demand in urban and rural populations



The funders need to recognise the differences that population density makes and adjust their approach to funding to encourage better capacity investment decisions. Funding growth for urban services can be relatively smooth, i.e., reflect demand growth minus a requirement for efficiency gains (including by making changes to better match the services to the population's needs).

On the other hand, because it is more difficult to match capacity and demand in more sparsely populated areas, the funders should be prepared to accept more variability over time in the provider's compliance with rural response time targets and be prepared to make lumpier increases to funding when response times fall below agreed levels for prolonged periods. Put another way, many rural areas should be able to absorb population growth and aging without adding capacity or needing increased funding for some time, but they will require a lumpy increase when new capacity is required. Applying the suggested urban approach to rural areas would encourage too much capacity investment too early in rural areas.

The consequence of adopting a more tailored approach to rural funding over the next few years needs to be considered in the context of the changes in the overall urban/rural funding mix suggested in section 6.3: Getting the rural funding base right.

6.3. Getting the rural funding base right

While current funding largely reflects the providers' existing costs in providing ERAS, it also happens to coincide with the distribution of funding from the Ministry and ACC that would result if both providers were funded solely on the basis of the size of their respective populations.

In terms of age distribution, the proportion of people over 65 years of age in each area is reasonably similar, i.e., in 2014, about 13.4% of the WFA population and 14.5% of the OSJ population were over 65 years old.

However, OSJ's population is significantly more rural than WFA's population. While any definition of 'rurality' will involve judgement, a reasonable, independent and consistent approach is to define the rural population for ERAS as those living more than 60-minutes' drive from a tertiary hospital. About

4% of the WFA population and about 35% of the OSJ population meet this criteria. OSJ also estimates that about 35% of its incidents are rural (which suggests that this hospital-based criteria defines rurality in a way that is recognisable and broadly consistent with the view of the providers).

WFA can service its very small rural population with double-crewed ambulances from the current per capita funding. OSJ has only managed to service its larger rural population from a similar per capita funding by single crewing much of its rural service. This is something that OSJ believes is no longer desirable, even when using multiple single-crewed vehicles to attend an incident (and including other first responders, such as the New Zealand Fire Service, to supplement).

6.3.1. Calculating the right loading for rural services

While it is currently likely to be more expensive per capita to service the rural population, whatever that current extra cost happens to be is covered by the current funding arrangements. The funders should consider an explicit and additional rural loading now if they accept OSJ's argument that it is no longer desirable to service its much larger rural population with a much cheaper, single-crewed, service.

OSJ believes it can address this situation with an additional 375 full-time equivalent (FTE) employees involving an additional cost of $22.5 million per annum. This $22.5 million translates into about 10.5% of the OSJ estimate of its cost of providing the ERAS, i.e., 10.5% of the $212 million estimated ERAS costs for 2015/16[4]. This suggests that it will cost at most about 30% more per capita to service the rural population relative to the non-rural population on the same double-crewed basis.[5]

The right rural weighting should be far less than 30% for three reasons.

1. The funders need to invest across ambulance and other services for the greatest return in terms of better prevention, treatment and rehabilitation outcomes. At some point, the annual reduction in single-crewed incidents becomes too low to justify the annual cost. On OSJ's plans, the cost of reducing a single-crewed response increases from about an additional $500 per single-crewed response avoided for the first 22 stations to about $1,000 for the next 29 stations to over an additional $2,000 per single-crewed response avoided for the last 45 stations.
2. In 'lower workload' stations, which receive fewer call-outs, providing for double crewing for all responses is not justifiable solely on the number of single-crewed responses reduced, but lower workload ambulance crews will have the capacity to provide a broader range of health services to the rural community. For these stations, the move to double crewing should be funded under the value-added category.
3. Part of the justification for returning to double crewing relates to the associated reduction in the dispatch of two single-crewed vehicles to the same incident. Savings from fewer double dispatches is not netted off the cost (though OSJ only expects the number of times two ambulances are sent to the same incident to fall from 5.9% to 4.1% with double crewing, so this saving is relatively small).

[4] Crown Funding Review. August 2015. Introduction, page 7. Paper provided by OSJ for this report's research.
[5] The total cost increases by 10.5%, and 35% of the population served is rural so 0.65 plus 0.35 times the increase in per capita cost of servicing the rural population is equal to 1.105 (0.65 + (0.35*1.3) = 1.105).

While the impact of this last factor is relatively small, the impact of the first two is significant, albeit a matter of judgement.

The funders must recognise that some rural loading of funding is justified. Rural populations will typically be more expensive to serve on a per capita basis because some capacity costs are fixed and there are economies of scale in station costs. While this has been absorbed to date in the overall funding quantum, the cost of double crewing means this is no longer possible for OSJ because of its large rural population. It would be better to explicitly recognise this as part of reducing single crewing, where that is deemed worthwhile, and apply a loading on the rural service evenly across the providers based on their rural population mix.

On the other hand, OSJ needs to accept that not all of its cost of comprehensive double crewing should be factored into this rural loading. Some of the cost will need to be justified and recovered as value-added services, i.e., the value that the extra ambulance personnel add to the local community in addition to their ERAS function. Some of the cost will not be able to be justified for ACC/Ministry funding because the savings in single-crewed responses is simply too low to justify the cost. Some will need to be spread over a longer time frame than currently envisaged by OSJ because ACC/Ministry capacity funding needs to be linked to the growth in capacity needed to meet demand. Demand growth will require the addition of new capacity at some point, and double crewing is likely to be the best way to meet that increased demand because it is a higher-value capacity and because, at that stage, it will also generate a bigger reduction in single-crewed responses.

6.3.2. A practical solution

A practical response would be to:

a. add an initial loading on rural funding paid to OSJ in the next three years that is sufficient to double crew areas where high demand is likely to require additional capacity in the short term anyway (because existing stations respond to more than 1,000 incidents a year) or where the reduction in single-crewed responses would justify the funders' investment. OSJ has prioritised double crewing into three categories on the basis of the expected reduction in single-crewed responses; with reductions over about 400 per annum initially (22 stations), then over 180 per annum (another 29 stations), then the rest (a further 45 stations). This corresponds to the increasing cost of reducing a single-crewed response referred to at point 1 in section 6.3.1: Calculating the right loading for rural services above (e.g., stations with reductions in single-crewed responses over 400 per annum have an average additional cost to reduce a single-crewed response of about $500)

b. apply that loading to the rural funding paid to WFA to cover its rural population in order to maintain funding parity and comparability between the providers

c. agree to phase in any further loading to fund rural services as the demand for rural capacity increases and/or the savings in single-crewed responses exceeds the level the funders can justify given other priorities.

So, for example, if the funders accepted the 400 per annum per station reduction in single-crewed responses as providing a reasonable initial return given other priorities, then a maximum of 40 OSJ rural stations would qualify for double crewing in the next three years on the above criteria. There are nine non-rural stations that would meet these criteria, and they should be picked up under the suggested approach to urban funding.

If we assume that each station in each of OSJ's three categories costs about the same to double crew, then the cost of double crewing these 40 rural stations would be about $10.6 million, which would suggest an initial rural premium of a little less than 15% (with that 15% spread over three years to the WFA rural population as well).

This may not guarantee double crewing in all situations as demand grows in these 40 stations. It may well be that extra vehicles are required to maintain response times as demand grows in the areas serviced by these stations, and some of these extra vehicles may need to be single crewed to meet the demand growth. However, if the number of single-crewed responses from any rural station starts to reach the agreed trigger level, say 400 responses per annum, then the funders should increase capacity funding to ensure single-crewed responses stay under the level.

6.4. Better matching response to need

The final element required to establish a regularised and consistent ERAS funding track is to align funding to the desire to better match provider response to patient need, given that one rationale for shifting to capacity funding was the recognition that not everyone was best served by an ambulance response that took them straight to hospital. This should also reduce the need to add as much new capacity as would otherwise have been required to meet the growth in demand.

6.4.1. Rural services

If something like the 15% rural loading described above is accepted, then most of the required new rural capacity for the next few years should be met through this mechanism (especially because the 40 OSJ rural stations would have been selected on the basis that they are most likely to need additional capacity, and they would be getting more staff). If we accept the arguments in preceding sections to determine rural funding, then any additional rural capacity funding would only be triggered if:

a. response times fell below current agreed levels across the whole rural population for prolonged periods (e.g., consistently for 12 months or more on a 12-month rolling basis);[6] or

b. if the number of single-crewed responses from a particular rural station increases beyond the agreed level that the funders see as giving them a reasonable return (say, 400 per annum).

The providers should be reporting their performance and processes to NASO to avoid hitting these trigger levels. Reaching these trigger levels should be seen as a reason to increase funding only if NASO agrees that all reasonable steps have been taken to avoid the triggers. In doing that, meeting the currently specified average response time across the rural population should take precedence over maintaining double-crewed responses. It is better for the patient to have a single-crewed ambulance arrive on time than to have a double-crewed ambulance arrive too late or not at all.

The providers report incidents closed by telephone advice (hear and treat), treated at the scene (see and treat), transported to a non-ED destination and transported to hospital ED, and national targets have been agreed separately for each provider based on current performance. The aim is to encourage more of the first three categories (especially the 'see and treat' category) and reduce the

[6] For context: current performance has been improving on a rolling 12-month basis and the providers typically meet their agreed response time targets.

number of transports to hospital EDs. The providers should be asked to provide separate data for their rural stations.

Given the suggested approach to funding rural services, the providers should make a concerted effort to extend existing capacity through these more tailored responses before seeking additional capacity. The best way to align capacity funding in rural areas to encourage these responses would be to make any additional investments in rural capacity conditional on minimal targets for at least the 'hear/see and treat' categories being met (albeit setting the 'hear and treat' target would depend on the national availability of the clinical hub). It may also make sense to include a 'transport to non-ED' target when that would reduce capacity utilised per incident response (e.g., because travel times are reduced) or when the DHB made a financial contribution to help achieve this outcome. Once a station has been double crewed, we would expect at least the 'see and treat' standard to increase.

6.4.2. Non-rural (urban) services

The suggested approach to funding urban capacity is quite different because capacity can be adjusted more smoothly to meet changes in demand, so funding demand is a good proxy for funding capacity.

Risk assignment

While the providers have little control over the volume of emergency demand growth, or economy-wide wage and price pressures, they have a significant influence over the cost of responding to this demand and over their organisation-specific costs. On the other hand, the funders are better placed to spread the risk of unexpectedly large increases in demand or economy-wide wage and cost pressures.

Rather than continuing to require the providers to absorb all of the costs associated with increased demand, the funders should meet these costs, minus an allowance for expected efficiency gains. This leaves the risk of stronger than forecast demand growth with the funders (who are best able to absorb it) while incentivising the providers to reduce the cost of meeting that demand (because they are best placed to manage this cost). This is similar to the logic for tying funding adjustments to an economy-wide measure of wage and price pressure, as happens now. The providers will need reassurance that method used to measure emergency demand growth will remain unchanged.

Setting the efficiency requirement

It is difficult to be precise about the quantum of efficiency gains that should be expected, although it seems unrealistic to continue with the current implicit assumption that OSJ can absorb all of what has been about 4% per annum growth. Efficiency gains of between 1.5–3% would be more realistic[7], so 2% per annum over a five-year period seems a reasonable starting point.

If we assume demand growth for OSJ remains at about 4% per annum, then the 2% efficiency gain assumption will leave half the cost of meeting this demand with the funders and half with OSJ.

[7] This reflects the variation in the New Zealand economy's measured sector labour productivity across various cycles, from about 1.5% (1982–1985; 2000–2008) to about 3% (1985–1990; 1997–2000). See Statistics New Zealand and New Zealand Treasury Productivity Paper 2008/02.

However, if demand grew at, say 5%, then the funders would end up funding 3% with the 2% efficiency gain assumption.

Measuring demand growth

This report suggests incorporating funding for communication centres into base funding so the providers face the true marginal cost and benefit of managing these centres. If adopted, this means that the cost of increased demand to the providers will become a mix of the cost of servicing both increasing emergency (111) calls and the incidents arising from those calls. However, if the suggested approach to communication centres is not adopted, then demand would be best measured by incidents alone.

The approach matters because the ratio of calls to incidents has increased significantly; especially as more people with mobile phones in the proximity of an incident are reporting the same incident and it is easier to make more follow-up calls for the same incident (e.g., to check ambulance progress or report changes in circumstances). Call growth would overstate incident growth.

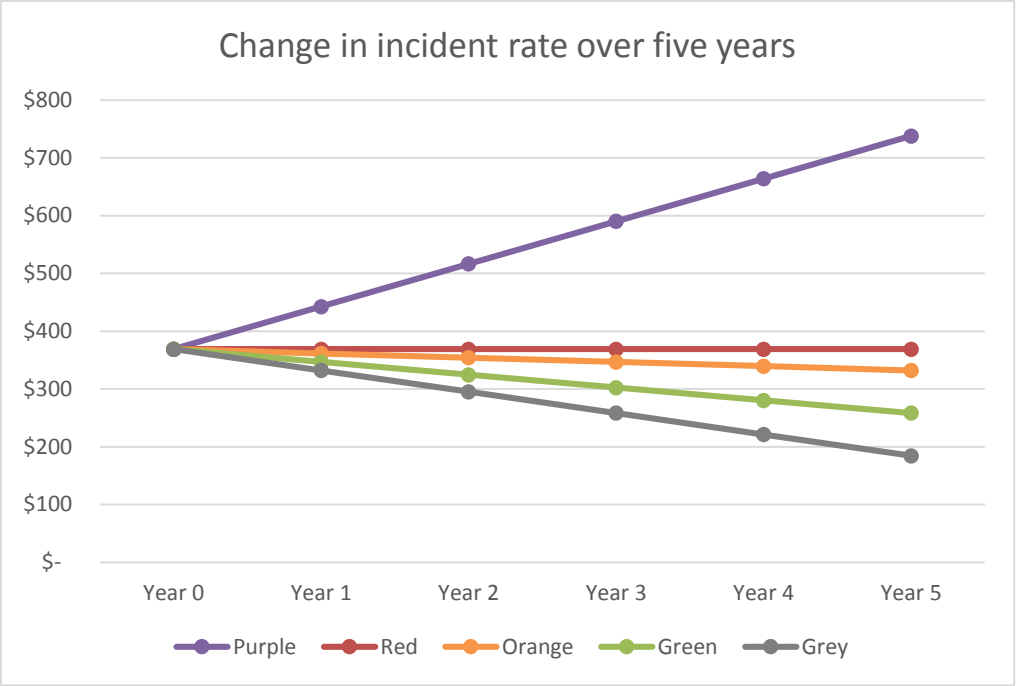
If the communication centre funding recommendation were adopted, then the right approach to measuring demand would be to combine both calls and incidents in a ratio that broadly reflects their relative cost. Because calls are cheaper to service, they should have a lower weight. Without good cost data, it is difficult to say what these weights should be. However, a generous position would be to start with calls as one-third of the weight attached to incidents. Thus, if responding to incidents were three times more expensive than responding to calls, then the formula for demand growth would be the sum of call growth (x 0.25) and incident growth (x 0.75). These weightings would need to be revisited within the first year of operation, when better information on relative cost was available.

Weighting incidents by priority to better align funding and value

While the funders could simply apply this relatively simple 'growth less efficiency gain' formula to the urban population, it does not reflect that fact that responding to higher priority incidents has more value to the population. Nor is it aligned to the changes in operating practice that the providers are trying to achieve.

An alternative is for the funding formula to give progressively more weight to the higher priority and more time-sensitive purple and red incidents and progressively less weight to the orange, green and grey incidents. This better aligns the funding to the providers' desire to more closely tailor different responses to different patients' needs because there should be more scope to substitute for transports (e.g., via 'hear and treat') in the less time-sensitive and typically lower acuity cases. Doing that will also reduce the capacity required – and so the cost to serve – orange, green and grey incidents. The funders will need to be assured that the assignment of priority ratings remains unchanged. Figure 2 illustrates the approach (for illustrative purposes only).

Figure 2: Change in incident rates over five years



The easiest way to operationalise this is to:

i. establish the base year by applying current funding evenly across the different priority 'colours' on the basis of the proportion of incidents that are prioritised to each colour (e.g., purple and red would earn a little under half of total funding)

ii. gradually increase the weighting given to purple (e.g., from a base weighting of 1 to a weighting of 2 over five years), hold red at the base weighting of 1 and reduce the weightings given to the other colours (e.g., reduce orange from 1 to 0.9, green from 1 to 0.75 and grey from 1 to 0.65 over the five-year period).

Total urban funding for each year would be the actual volume of incidents in each of the priority colour groups for that year weighted by the $ weights described in step (ii) above.

The final funding for volume growth will be a function of the weightings adopted in step (ii) above. The weights used as examples here are consistent with efficiency gains of about 2% per annum over the five years. Faster (slower) declines in the weighting of lower priority incidents are consistent with larger (smaller) assumed efficiency gains.

This approach creates a strong incentive on the providers to shift as many calls and incidents as possible into the 111 priority rating system and to give each incident the highest rating that management discretion allows. Given the large number of calls that are not currently prioritised in this way, this could lead to substantial re-categorisation. While the impact of this incentive may be modified by the impact on recorded response times for higher priority incidents, the funders should not rely on this being sufficient. If the providers want the certainty that comes with tying funding to demand growth, they need to assure the funders that this risk can be contained.

At least three reinforcing approaches should be required to help mitigate the risk of providers elevating the priority rating for incidents. The funders should agree on the protocol that the providers will use to include a call and associated incident in the priority rating scheme and then

assign a priority to that incident, with this protocol reflecting current practice. Compliance with those protocols should be audited annually by one of the large audit firms.

Finally, NASO should monitor the growth in incidents and the priority mix of incidents and compare that to predicted growth given increases in the age-adjusted population. Any significant divergence should trigger a review of the way the providers are applying the approach and possibly of the approach itself.

These would seem to be the minimum steps required to ensure that what will be a very new process is seen to have integrity and, therefore, more likely to have durability.

Population-based urban funding

An alternative approach would be to tie urban funding increases to changes in age-adjusted population. However, this introduces the risk that growth in age-adjusted population will not pick up all the influences on demand or that it will not weigh these influences correctly. Both risks undermine funding certainty because ultimately both the funders and the providers will come back to actual growth in incident demand. It would be better to focus on incident demand and monitor the relationship between that and what would be expected, as suggested above, to confirm that demand is being recorded correctly.

6.5. Summary: Implications for constructing an ERAS funding track

The proposed way forward treats the rural track quite differently from the urban track, for the reasons explained above.

6.5.1. The proposed urban ERAS funding track

a. Current funding sets the base year amount for ACC and Ministry contributions. This base was established after a thorough review of provider cost models, albeit some re-basing of ACC contribution based on actual costs was anticipated. The urban component would be 96% of WFA and 65% of OSJ base funding from ACC and the Ministry.

b. Each year, the previous year's funding would be increased by a combination of:

    i. the current cost adjuster to compensate for economy-wide inflation; and

    ii. real increases in annual funding, which would be determined by applying the growth measure and the priority-weighted formula described in section 6.4.2 (i) and (ii)

    iii. plus or minus the amount needed to offset any decrease (increase) in required urban response times.

6.5.2. The proposed rural ERAS funding track

a. If the assumptions made in section 6.3: Getting the rural funding base right were accepted, then an initial 15% loading would be added to the 'premium' for the rural proportion of the covered population. In OSJ's case, about $10.6 million would be added to its funding over the next three years on the condition that this funding be applied to implement the planned staff increase necessary to ensure double crewing of the 40 OSJ rural stations that meet the criteria in that section. If different criteria were applied, then this rural premium would be increased or decreased proportionately. The same 15% premium would be added to the 4%

rural population covered by WFA (i.e., an increase in funding for this 15% component of 4% of the total spread over three years).

b. The annual cost adjuster would remain for rural services.

c. Funding would increase (decrease) in proportion to any decrease (increase) in rural response times required by the funders in future.

d. OSJ should be required to report individual rural station performance in terms of single-crewed responses.

e. Any additional capacity funding would only be triggered if the providers had met agreed minimal targets for the proportion of rural calls managed on a 'hear/see and treat' basis **and**:

   - rural response time performance fell below currently specified target levels for prolonged periods (e.g., consistently for 12 months or more on a 12-month rolling basis across the whole rural population); **or**

   - if the number of single-crew responses increases beyond the agreed level that the funders see as giving them a reasonable return (say, 400 per annum on an individual station basis); **and**

   - NASO agreed that the providers had taken all reasonable steps to avoid reaching these trigger levels, with meeting agreed response times taking precedence over double-crewed responses.

f. Any further increase in the rural loading would only occur as growth in demand raised the expected savings in single-crewed incidents above a level determined by the funders to represent a good return on investment relative to other health, preventative and rehabilitative investments (e.g., the 400-per annum level from any particular station).

7. Funding value-added services

The transformation in OSJ's and WFA's business models creates an opportunity for these ambulance services to provide a much wider range of preventative, rehabilitative and curative services in addition to their core ERAS. These have the potential to add value by improving outcomes and saving cost throughout the health system. This potential additional value falls into three broad categories:

- Improved patient outcomes, i.e., survival and better/faster return to function than expected under current arrangements by implementing the ambulance components of new health system clinical protocols and pathways. WFA and OSJ are already in discussions with the relevant clinical networks (trauma, stroke, cardiac arrest, etc).
- Reduced occurrence/severity of injury/illness than expected under current arrangements. There is already some focus on key areas for injury prevention funded directly by ACC (e.g., falls in the elderly). Ambulance services are also well placed to provide advanced health monitoring and response, which allows people to self-manage at home for longer.
- Reduced hospital costs via reduced avoidable admissions and readmissions. The ambulance service is already making significant gains in this area and a trial funded by the Ministry demonstrated that OSJ could make significant additional reductions in avoidable admissions to hospital ED by introducing the model nationwide.

7.1. Improved patient outcomes and reduced occurrence/severity of injury/illness

Some of these services will be of direct interest to ACC and the Ministry, and these could be funded directly by them on top of the core ERAS funding on the basis of the value of each additional service provided (i.e., on the basis of better outcomes to patients, such as faster return to function).

This is relatively straightforward in the case of ACC, who has a well-developed framework for assessing actuarial benefits from improved outcomes and a strong incentive to invest in reducing future claims. Moreover, ACC directly funds services that provide these benefits.

The health sector currently lacks this sort of outcome framework so has weaker incentives to fund services that would contribute to the results described under the first two bullet points above. While changes that would address this issue are being considered, including the development of system-level outcome measures, for the purposes of this report, the assumption is that current health funding arrangements will remain unchanged.

While the Ministry funds a small proportion of health services directly, these are unlikely to be services where the ambulance service can add a lot of additional value. Most of the benefits that the new ambulance business model can deliver would accrue to one or more of 20 DHBs or more numerous PHOs and may require their active participation for the benefits to be realised.

Helping PHOs provide quality after-hours care, especially in rural areas, is an obvious area for closer collaboration between PHOs and the ambulance service. From January 2016, OSJ has been a member of the DHB/PHO Rural Advisory Group. It would make sense for the Ministry to ask this group to recommend how best to configure services to sustain quality after-hours primary health care in rural areas. If necessary, the Ministry could then reflect these suggestions in its annual planning guidelines to DHBs and/or PHO agreement negotiations.

The Ministry also advises government on the targets that DHBs should aim to meet and monitors DHB performance against those targets. If the ambulance service can convince the relevant clinical networks that they have a role to play in improving outcomes for patients suffering trauma, stroke, cardiac arrest, etc and the resulting protocols are funded, then the ambulance service needs to be fully funded for the additional costs they incur. These costs should be identified by the relevant clinical networks before new protocols are agreed and funded.

Ideally, DHBs would fund the relevant ambulance service for the role they are expected to play. However, given current DHB funding incentives and the transactions cost of dealing with each DHB separately, this may not be practicable, and so the Ministry may have to fund the ambulance service and DHBs separately once new protocols are agreed.

7.2. Reduced hospital costs via reduced avoidable admissions and readmissions

A different approach would be required to fund ambulance providers in reducing avoidable admission and readmission to hospitals (including by supporting better home-based care, especially for people with chronic conditions, as well as reducing transports to hospital EDs).

The only benefit that ambulance providers gain from reducing transports to hospital EDs is some transport time (alternative destinations are typically closer). By far the largest gains accrue to the DHBs (according to a 2014 report by the Sapere Research Group, 30% of hospital ED presentations arrive by ambulance[8]).

DHBs need to be actively involved in order to realise the gains from reduced avoidable hospital admissions and readmissions. Ideally, they would be engaged in identifying where better ambulance practice is likely to generate the most savings – and for which DHBs – and helping to fund ambulances to deliver these gains.

Engaging all DHBs individually is likely to be a long and costly process. DHBs are busy, and many will assume that any changes will have only a marginal impact on their costs, or that the DHB will benefit from reduced avoidable hospital admissions and readmissions whether they are involved or not. While these assumptions may well prove correct in some cases, they are not the recipe for active engagement.

It makes more sense to engage selectively with those DHBs who will benefit most from additional ambulance services. These DHBs could be identified through a variety of performance metrics (e.g., ED wait times, acute medical bed days, readmission rates, acute length of stay). When additional ambulance services are likely to help lift performance, the Ministry could require the DHBs to include OSJ or WFA in the development of their annual plans. The *2016/17 Annual Planning Guidelines*[9] already asks DHBs to engage with 'relevant stakeholders' and agree on who should do what to achieve annual plan priorities.

The Ministry should also make it clear to those DHBs most likely to make capital bids in the next two to three years that securing Ministry support for additional capital investment in expanding ED and related capacity will be conditional on that DHBs effective previous engagement with the relevant ambulance provider. That engagement should be aimed at ensuring that the ambulance service is

[8] Sapere Research Group. April 2014. Estimate of the value to be generated by St John's new service model.
[9] See: http://nsfl.health.govt.nz/dhb-planning-package/201617-planning-package-and-review-plans/annual-plan-guidelines

being used to its full economic potential in helping reduce avoidable hospital admissions and readmissions. Avoiding investment in new capacity is the time when the bankable gains from reducing hospital admissions and readmissions are likely to be greatest.

8. Funding communication centres

The Terms of Reference ask 'At what point should there be greater collaboration between OSJ and WFA (i.e., through sharing of functions, overheads and infrastructure)?'

The providers were reluctant to explore these options because of the potential implications for their managerial and organisational autonomy, so it is difficult to know the degree to which obvious savings would come at the expense of operational effectiveness and, therefore, service quality. It would be far better to keep this option open for consideration as part of the approach to conditions of funding, discussed previously in section 5.3.2: Funding conditions.

Having said that, the communication centres probably represent the largest single area for savings from consolidation. Moreover, the communications infrastructure is aging, and major investment is likely to be required within the next three years. Therefore, it is particularly important that the providers are presented with the right incentives to make the best investment decisions in a timely way.

Current funding arrangements do not encourage this outcome. The providers' communication centres are funded separately from ERAS. Funders meet the total cost less a fixed $2 million annual combined contribution from the two providers. A fixed provider contribution does not create the right incentive for the providers to contain costs or rationalise these communication centres because the benefits of these decisions do not automatically flow to them.

It is particularly important that the providers face the true marginal cost of their decisions as they approach choices about infrastructure replacement. This can be achieved by reversing the current arrangement, i.e., giving each provider their share of the current Ministry and ACC funding for communication centres as the baseline funding amount from government for ambulance communication services, letting them fund the rest and discontinuing separate funding for their communication services. The current funding would be added into each provider's base ERAS funding and, therefore, be subject to the annual cost and volume-based increases recommended above (so this additional future funding stream provides part of the basis for financing communications asset replacement).

Carrying over the current levels of funder support for the communication centres probably overstates the contribution that they should make to support ERAS because they support non-ERAS as well (e.g., alarms, PTSs, etc). However, this is a small consideration compared with ensuring that the providers face the true marginal cost of the decisions they take.

It may well be, for example, that WFA would be far better placed if OSJ owned and operated all the communication centre capacity and WFA simply paid to use that capacity. That could generate $1.5–$2 million in savings that could be shared between the two providers. However, WFA is unlikely to reach this conclusion if the benefits of so doing accrue to the funders and WFA has to compromise some organisational autonomy.

9. Feedback from the providers

The author presented the approach and recommendations recorded in this report to OSJ and WFA Boards and left each with a copy of his presentation notes (including the initial draft recommendations). Both organisations provided feedback, which is summarised below.

9.1. St John

OSJ saw the proposals as 'a significantly better basis' to support ERAS than the current approach. It fully supported the main elements of the suggested approach: funding conditions; the overall approach to funding a five-year trajectory based on actual demand; targeted funding to reduce single crewing (albeit more targeted than OSJ originally proposed); the proposed urban funding model (although it considers a 1–1.5% efficiency dividend to be more realistic than the 2% suggested in the report); the proposed rural funding model (but introduced over two rather than three years and starting in 2016/17); introducing a national PTS contract based on full cost recovery; and incorporating communication centre funding within the base EAS funding (albeit with an initial distribution between OSJ and WFA based on an increased base-funding level).

However, OSJ believes that there are two conditions that need to be met before the new funding arrangements are applied. These are as follows.

- Its funding base should be increased to incorporate '… the agreed roll-forward of ACC capacity funding levels from the five-year period 2012/13 to 2016/17' (worth about $4.8 million if applied to 30 June 2016 on OSJ's calculations).
- The annual cost adjustment should be linked to objective national indices, rather than have what OSJ argues is an unpredictable discretionary element applied to it.

OSJ also thought that further development was required in relation to: introduction of priority-weighting (suggesting this be 'shadow priced' and introduced after two to three years' experience and assessment); using financial incentives to encourage ARC providers to economise on ambulance use (preferring the Ministry to support the introduction of an agreed ARC transport protocol combined with some focus on ARC facilities with high ambulance utilisation rates); complementing the suggested approach to value added services by resourcing NASO to better 'represent ambulance sector interests across the health system' along with monthly meetings between OSJ, WFA and the heads of ACC and the Ministry of Health to 'co-ordinate value-based business case submissions'.

9.1.1. Commentary on this response

OSJ's main concern is to secure a funding adjustment in 2016/17 to increase its funding base in recognition of the substantial real demand growth it has absorbed since the 2012/13 base year and to be able to initiate its programme of double crewing.

Given the reality of the demand growth that OSJ has absorbed since 2012/13 and the expected ACC roll-forward, there is a reasonable case for this $4.8 million re-basing before the new funding approach is introduced.[10] Given the desire from both the funders and the providers for greater

[10] While the author has not been able to cite a written commitment from ACC to re-base its funding to 40% of actual ERAS costs, it is understood that this was the intention communicated to OSJ and WFA at the time when the current funding arrangement was introduced in 2014.

certainty, OSJ's suggestion that the annual cost adjustment be made on the basis of objective national indices, such as a blend of the labour cost, consumers price and producers price indices (LCI, CPI and PPI respectively), also makes sense.[11] This would meet both the conditions that OSJ suggests be met before the new funding approach is introduced.

A logical way forward would be to re-base in 2016/17 and introduce the new funding approach in 2017/18 on the new base.

However, OSJ would also like to see a 2016/17 start to adding the rural loading so that it can initiate double crewing in the highest priority rural areas this year.

Either of these changes would represent a significant extra funding commitment in 2016/17 at a time when OSJ is forecasting sizable deficits. If they were to be considered favourably by the funders, then OSJ's forecast deficits would trigger the conditionality element of the new arrangements. Unless OSJ could convince the funders that a $4.8 million re-basing plus any 'stage 1' corrective actions contemplated for 2016/17 were sufficient to address the forecast deficit, then the funders should contemplate moving directly to 'stage 2' conditions, i.e., requiring their approval of OSJ's budget before the budget is finally agreed by the OSJ Board.

OSJ's desire to reduce the required efficiency dividend from 2% to 'something closer to 1–1.5%' is based on the observation that this dividend has to be secured against 100% of its costs rather than just the 70% funded by ACC and the Ministry. While true, the 2% suggested in this report was based on a 1.5–3% range that was the range for labour productivity seen across the whole measured sector in the economy and so applies at the enterprise level, i.e., on 100% of an enterprise's entire cost base. While the report acknowledges this is an area for judgement, it is hard to see why efficiency gains expected of ambulance providers should be less than that achieved on average by other New Zealand enterprises.

OSJ's actual annual demand growth has varied over the last three years, so if the parties eventually agreed to 1.5%, the funders would be carrying a significant risk in high growth years. Given the providers have some scope to smooth the impact of strong growth in any one year, a compromise solution might be to reduce the dividend to 1.5% but either:

- cap any single year's funding contribution to say 4%; with any growth in excess of the 4% added to subsequent years funding. So an 8% year followed by a 2% year would be funded with a 4% then 3% increase rather than a 6.5% and 0.5% increase respectively; or
- use a two-year moving average in demand growth rather than annual demand growth (for a financial year ending 30 June, the winter peak can fall between June and September, so it may fall twice in the same year).

OSJ's other issues are typically of a lower order of concern.

- Given the newness of the recommended priority weighted approach to urban growth, a period of 'shadow pricing' may have merit. While the *expected* funding quantum of a weighted and unweighted approach is the same, *actual* funding would be higher/lower if the priority mix of actual incidents were higher/lower than expected. The unweighted approach

[11] While there is some dispute about the degree of discretion applied in practice, formal agreement to a non-discretionary approach will remove future uncertainty.

is probably easier to understand and does not create a financial incentive to reweight incidents, so would involve less auditing. On the other hand, funding is not so clearly aligned with the value provided by the service and the changes in operating practice that the ambulance providers are trying to achieve. While the unweighted approach could be complemented by adding specific targets for alternative responses to low priority cases (e.g., more 'see and treat' beyond purple and red responses), this adds a complication to the unweighted approach and might prove harder to justify and enforce if targets were not met.

- OSJ's suggested approach to ARC would not create as strong incentives on ARC facilities to use ambulance services wisely as financial incentives but would deliver some improvement (and probably with less opposition).

- OSJ's suggestion that NASO be resourced to better represent ambulance providers to the rest of the health sector would significantly change NASO's role and is not recommended. Adding regular meetings with senior Ministry and ACC officials to coordinate value-added business case submissions would be helpful in some cases but is no substitute for direct engagement with DHBs, as suggested in this report. Direct engagement with ACC to provide value-adding services to ACC makes sense. However, while the Ministry can help encourage the right outcome in health, it is the DHBs who will need to fund ambulance services that either add value to the DHB population or reduce DHB cost.

9.2. Wellington Free Ambulance

The funding changes suggested in this report are far less valuable to WFA for four main reasons:

1. Ambulance demand growth has been relatively flat and is likely to remain subdued in the WFA area (so there may be little, if any, increase in funding from real emergency demand growth; especially after any required efficiency gains are deducted).
2. Only 4% of the WFA population is rural so the funding benefit from the proposed rural loading applied equally to OSJ and WFA is relatively modest for WFA.
3. As the name suggests, WFA does not impose part charges and so increasing flexibility in this areas has no financial benefit to WFA (although this is a business decision that WFA makes).
4. WFA does not believe that it can secure extra funding from other sources.

Given this, it is understandable that WFA's main concerns are around:

- the adequacy of the original 2012 funding base, and

- a desire to see government fund a higher proportion of WFA's costs.

In regard to the first concern, WFA raises the same point as OSJ regarding the intention to re-assess total service costs in 2016/17 (i.e., the cost of current capacity) and ACC's commitment to funding 40% of the re-assessed cost. The two providers should be treated equally in this regard in terms of re-basing funding.

The argument for equal treatment should also be applied to the second concern. It is, therefore, hard to see why WFA should receive more government funding than OSJ on a per capita basis (adjusted for rurality and increased demand over time) for providing its population with emergency road ambulance services.

9.2.1. Commentary – funding conditions

WFA argues that it should only be 70% accountable to the funders on the basis that they only fund 70% of WFA's cost. This goes to the heart of the moral hazard problem. This report recognises the value of organisational autonomy and, in section 5.1: Provider autonomy is desirable and requires financial flexibility, that 'the quid pro quo of this autonomy is that OSJ and WFA are able to live within their means, i.e., that they do not come back to the Ministry and ACC to fund poor decisions or cover financial risks that have not been well managed.'

While WFA can maintain the financial capacity to manage these risks itself, ACC and the Ministry would not attach conditions to the funding they provide. However, the recommendation is that conditions be imposed requiring WFA to '… take specific actions to improve its financial position, without recourse to the funders, as its ability to manage financial risk deteriorates' (see section 5.3: Accountability arrangements need to address this co-dependence). These conditions must be able to cover all of WFA's operations in order to ensure that steps can be taken to improve the financial position *without recourse to the funders*.

The suggestion that WFA should maintain autonomy over 30% of its revenue – and, presumably, some or all of its assets – at a time when it needs to come back to the funders for more money means that WFA would be looking to the funders alone, rather than the full range of options available, to improve its financial position. This would leave ACC and the Ministry in the unenviable position of being the funders of 'first resort', having to underwrite any omissions or poor decisions that WFA might make, which undermines WFA's need to take care and manage its risks well.

It is hard to believe that ACC and the Ministry would not require WFA to take remedial action across the full spectrum of its expenditure and asset choices if WFA needed these funders to bail it out of financial difficulty, even under current arrangements. However, the difficulty with the current arrangements is that this is not explicit, as WFA's response indicates, which weakens the incentive for WFA to manage its budgets carefully.

The advantage of the proposed arrangement is that it addresses the moral hazard issue upfront so WFA is clear that the funders will be expecting it to exercise the full range of remedial actions before it can come back to its funders for more money and, at that point, conditions will be required across this full spectrum before any additional funding is considered.

9.2.2. Commentary – funding demand and cost pressure

WFA argues that the demand it faces includes the increasing complexity as well as volume of demand: complexity of patients and cases ('not only of medical but also social needs') and in terms of delivering a wider range of services to this population (increasing numbers of patients seen and treated at home and involvement in treatment pathways for conditions like, such as heart attacks, STEMI, spinal injuries, etc.).

This report makes a distinction between ERAS capacity funding and additional funding for value-added services, and it is in this latter category that funding for the sort of complexity that WFA highlights should be considered. The report recommends that these be funded on the basis of the value added by the additional service provided (and assessed as part of a business case and against competing priorities that the funders have in order to improve health and rehabilitative outcomes). The alternative is to fund whatever costs ambulance providers decide they should incur as they expand their role in responding to this complexity. That not only frustrates the funders' ability to set

funding priorities, which is fundamentally a funding decision, but also encourages other health providers to shift costs to ambulance.

Current funding of WFA and OSJ is virtually the same on an age-adjusted per capita basis, and the proposed funding formula simply adds a 15% rural loading to reflect the extra value of double crewing in rural areas. WFA points out that it has already invested in double crewing. However, WFA will not miss out on the extra funding the report is suggesting for this activity. This loading is applied equally to both providers on the basis of the relative size of their rural populations (so compensating WFA for the double crewing it is already doing for its smaller rural population).

Like OSJ, WFA has concerns about the requirement for a 2% efficiency gain in calculating funding for urban emergency call volume growth. While WFA claims that 'increased demand eats into efficiency gains' actual emergency volume growth in the WFA region has been essentially flat. While demand remains subdued, the difference between the 2% suggested in the report and the 1–1.5% suggested by OSJ is largely irrelevant to WFA's financial position.

WFA argues that it 'would need an annual cost adjuster as a minimum', and maintenance of this adjuster is recommended in the report. WFA would also benefit if the funders adopted OSJ's suggestion to base this cost adjuster on objective national indices.

Like OSJ, WFA is also reluctant to accept the priority weighted approach to funding emergency demand and sees it as 'too complex'. The pros and cons of adding this feature are discussed in the OSJ commentary above, although it is easy to understand why the providers are reluctant to add complexity when so many other elements of funding would be changing. Given this reluctance, alongside the added monitoring cost, it makes sense to have a trial period before introducing priority weights.

9.2.3. Commentary – linking funding to response time requirements

WFA has misinterpreted the recommendations on linking funding to 'required' response times (see paragraphs 2.1.3 and 2.2.3 of section 10.2: Funding conditions and funding demand). This report is not suggesting any change to the current arrangements (other than allowing more flexibility in rural areas). The sole purpose of these recommendations is to recognise that more/less funding would be needed if the response time *required* by the funders was to become more/less demanding in future. The recommendations have been redrafted to make this clearer.

9.2.4. Commentary – funding from other sources

WFA does not see any opportunity for increased funding from other parts of the heath sector, such as DHBs, unless the Ministry specifically adds extra funding to the sector for this purpose. The assumption is that any improvement ambulance services might make to the rest of the health service will be captured by other health providers and not returned to ambulance providers. While this is a risk, this report suggests how the Ministry might address this risk in the way it deals with these other providers.

9.2.5. Commentary – funding communication centres

While WFA does not comment on the recommendations for funding communication centres, it does suggest increasing the funding of these centres.

10. Conclusions and recommendations

The feedback from the providers has resulted in modifications to the recommendations in the body of this report, including alternative options listed in the recommendations below.

10.1. Re-basing

The main concern of both OSJ and WFA is to re-set their base level of funding before the new funding arrangements apply (based on their expectation that ACC would fund 40% of actual ERAS costs). While this report has been focussed on recommending a better future funding arrangement for ambulance services, the funders may well have to address this re-basing issue in order to move forward.

If the funders agree to this re-basing, then these costs could be reviewed in 2016/17 and the new baseline applied with effect from 1 July 2017 on the proviso that:

- if the funding approach recommended in this report is also adopted, then there should be no expectation of any future funding adjustments based on the providers' costs. If the providers' costs exceed the available funding under the new arrangements, then the providers would have to address their costs, asset base and/or pricing decisions rather than look to ACC and the Ministry for extra funding; and
- given OSJ's forecast deficits, funding conditionality would be imposed at either the 'stage 1' or 'stage 2' (i.e., budget approval) level depending on the degree to which remedial actions on stage 1 were insufficient to both address these forecast deficits and ensure enough financial flexibility to absorb normal financial risk.

10.2. Funding conditions and funding demand

The recommendations on funding conditions and funding demand should be considered together and seen as inseparable.

Conditions for funding help counter the moral hazard implicit in the degree of co-dependency in the funder-provider relationship. Automatically funding some demand growth without addressing the moral hazard issue increases the cost of this moral hazard to the funders because it allows the providers to spend more before having to ask for more (which would otherwise be the point at which the funders might impose conditions). However, funding conditions will not create the right incentives for the providers if a deteriorating financial position is inevitable, which is likely unless some proportion of demand growth is funded on a predictable basis.

The timing of capacity investment is one of the most important decisions that a provider will make: investing too early increases cost unnecessarily, while investing too late risks degrading response times. The more densely populated an area, the greater the ability to spread lumpy investments over a larger population base, and so funding capacity growth in urban areas can be approximated by funding some proportion of demand growth. While the providers have little control over the volume of emergency service demand growth, they have a significant influence over the cost of responding to this demand. The funders are better placed to spread the risk of unexpectedly large increases in emergency service demand. Rather than continuing to require the providers to absorb all of these costs, the funders should meet these costs, minus an allowance for expected efficiency gains.

For rural areas, it will be impossible to match capacity to emergency service demand in the short and possibly even medium term, and so a different approach is needed to encourage the right timing of capacity investment. At the same time, adding a rural loading for the rural population served by each provider recognises the desirability of a significant reduction in single crewing in rural areas and allows this to be applied equally to both providers.

The two main recommendations in this report are that:

1. A scheme of conditions for funding be introduced as described in section 5.3: Accountability arrangements need to address co-dependence based on:

    1.1. disclosure and a set of parameters that will trigger a requirement for operational or financial management changes by the provider (i.e. a combination of some or all of the actions described in section 5.3, depending on the circumstances); and

    1.2. if those actions proved insufficient to restore financial flexibility, then that would trigger a requirement that the provider needs to have its budget approved by the funder until financial flexibility was restored or a change in management or ownership became inevitable.

2. The funders increase funding each year.

    2.1. For **urban** populations (i.e., for 96% of the WFA funding base and 65% of the OSJ funding base), funding is subsequently increased by:

        2.1.1. an annual adjustment to recognise economy-wide cost increases that are non-discretionary and based on objective national indices (e.g., a weighted combination of the LCI, CPI and/or PPI)

        2.1.2. a real increase each year based on the annual percentage increase of either:

            a. the weighted sum of emergency call growth and incident growth, with the weights based on the relative cost of calls and incidents and agreed before 1 July 2017 (e.g., if responding to incidents were three times more expensive than responding to calls, then the formula would be: emergency call growth (x 0.25) plus incident growth (x 0.75); or

            b. incident growth alone (if the communication centre recommendations below are not accepted)

            less an efficiency dividend of either:

            a. 2% per annum; or

            b. 1.5% per annum with a cap on the funders' contribution of 4% in any one year with the funding for growth in excess of that cap spread over subsequent years

            calculated for the first two years on the basis of demand that is not weighted by priority while the impact of applying the priority-weighted approach described in section 6.4.2: Non-rural (urban) services is assessed

2.1.3. plus or minus an amount for the cost impact of any decrease (increase) in urban response times required by the funders.[12]

2.2. For **rural** populations:

2.2.1. an initial 15% rural loading is added to 4% of WFA's funding base and 35% of OSJ's funding base (being about $10.6 million) on the condition that OSJ's loading is used to double crew the 40 rural stations that meet the criteria set out in section 6.3.2: A practical solution and is spread over three years, starting in either 2016/17 or 2017/18

2.2.2. an annual adjustment is made to recognise economy-wide cost increases that is non-discretionary and based on objective national indices (e.g., a weighted combination of the LCI, CPI and/or PPI)

2.2.3. plus or minus an amount for the cost impact of any decrease (increase) in rural response times required by the funders

2.2.4. OSJ should be required to report individual rural station performance in terms of numbers and percentages of single-crewed responses

2.2.5. any additional capacity funding would only be triggered if the providers had met agreed minimum targets for the proportion of rural calls managed on a 'see and treat' and 'hear and treat' basis **and**:

- rural response time performance falls below currently specified target levels for prolonged periods (e.g., consistently for 12 months or more on a 12-month rolling basis across the whole rural population); **or**
- the number of single crew responses increases beyond the agreed level that the funders see as giving them a reasonable return (say, 400 per annum on an individual station basis); **and**
- NASO has agreed that the providers took all reasonable steps to avoid reaching these trigger levels, with meeting agreed response times taking precedence over double-crewed responses

2.2.6. any further increase in loading for rural responses would only occur as growth in demand raised the expected savings in single-crewed incidents above a level determined by the funders to represent a good return on investment relative to other health, preventative and rehabilitative investments (e.g., the 400 per annum level from any particular station).

These annual increases could apply from 1 July 2016 (without re-basing) or from 1 July 2017 (on the re-based level) or a combination of the two approaches.

[12] The change in costs and funding required to meet a change in response-time requirements would be negotiated.

10.3. Increasing third-party funding for core ERAS

The Ministry should support the introduction of an agreed ARC transport protocol combined with some focus on ARC facilities with high ambulance utilisation rates with the aim of reducing those rates. In the meantime, NASO should also identify the actual quantum of demand that is generated by ARC facilities and the range of actions the Ministry might take to improve the efficiency of ambulance utilisation.

The Ministry should increase the maximum part charge that ambulance providers are able to charge.

The Ministry should require DHBs to pay the full cost of PTSs provided by the two ambulance providers (perhaps as part of a national contracting approach).

10.4. Funding value-added services

The two ambulance providers are wanting to add a wider range of care, treatment, advice and referral services to their core emergency response and transport service. Demand for these wider services is additional to emergency ambulance demand. Funding these wider range of services needs to be considered separately from funding the emergency road ambulance service and be assessed alongside other funding priorities. Continuing to fund ambulance providers on the basis of the costs they incur as they expand their role would effectively pre-empt the funders ability to set funding priorities and encourage other health providers to shift cost to ambulance.

This wider range of services potentially adds value to DHBs and PHOs as well as the Ministry and ACC and could be funded in a number of ways.

Relevant clinical networks are identifying the role that the ambulance providers should play in each of the speciality areas where treatment protocols are currently being redesigned to improve patient outcomes. They should be asked to work with ambulance services to identify the cost to the ambulance services of playing this role. Once the funders have agreed that the new protocols are cost-effective and will be funded, then the Ministry should ensure that ambulance providers are fully compensated for the extra costs they incur in supporting these improved outcomes.

The Ministry should ask the Rural Advisory Group to recommend how best to use ambulance in helping PHOs meet their after-hours service obligations in rural areas and reflect these suggestions in DHB Annual Plan Guidelines and/or the PHO Services Agreement.

The Ministry should ensure that DHBs have engaged ambulance providers in the development of their annual plans and are funding additional ambulances services where this is the best way to deliver the priorities in their plans.

The Ministry should also make it clear to DHBs that securing Ministry support for additional capital investment in expanding ED and related hospital capacity will be conditional on effective engagement with the relevant ambulance provider (so the ambulance service is funded to ensure that it is helping to reduce avoidable hospital admissions and readmissions when that makes economic sense). The Ministry should also make it clear that it expects DHBs to reward ambulance providers for reducing avoidable admissions and readmissions where they are best placed to do so economically (rather than effectively shifting the cost of this onto the ERAS).

ACC and the Ministry should consider OSJ's suggestion of regular high-level meetings to coordinate value-added business case proposals.

10.5. Funding communication centres

In order to ensure that the providers face the true marginal cost of owning and operating the communication centres, the funding that the Ministry and ACC provide to support the communication centres should be fixed at the current level and allocated proportionately to WFA and OSJ as part of a single ERAS funding stream. Separate funding of these centres should cease.

The proportion of this fixed amount currently allocated to each provider should be added to their base funding and so subject to the annual index-based cost and volume-adjusted increases recommended above. That should provide a stream of future revenue that can be used to finance replacement of communications assets, when required. If this is not sufficient to maintain three centres, as implied by WFA's comments, then consideration should be given to consolidation into two centres.

10.6. Implications for NASO

While NASO should be able to manage this regime when funding is unconditional, as is currently the case, it may need supplementary resourcing if conditions for funding are triggered – especially if either provider moves into the territory that requires the funders to approve their budgets.

Appendix 1: Terms of reference

Review of the funding arrangements for emergency road ambulance services, including ambulance communications

Purpose

The review will provide the Director General of the Ministry of Health (Ministry) and the Chief Executive of the Accident Compensation Corporation (ACC) with advice on government funding arrangements that will provide quality, innovative, and sustainable emergency road ambulance services.

Background

In recent years ambulance services have continued to transition from being an emergency response and transport service to becoming highly-qualified urgent healthcare providers, connected with the wider health system, providing any combination of care, treatment, advice, and referral appropriate for each individual.

Emergency ambulance services today close 3% of 111 calls by providing advice by telephone, provide care at the scene without needing to transport in 25% of calls, and transport about 72% of calls to hospital. They are on a pathway to (a) increase the proportions of those closed by telephone advice and those managed at the scene without transport, and (b) reduce transports to hospital to 55% of all calls by 2020.[13]

As well as changing the way they work, ambulance services face increasing demand. The number of 111 ambulance calls is increasing at 4% per annum and is expected to grow faster over the next several years due to demographic and socio-economic drivers. While ambulance services can influence the cost of servicing these 111 calls, they cannot offset the full cost of increased volume and continue to improve service quality.

Emergency ambulance services are primarily funded by the Ministry and ACC, which together cover around 70% of total costs. The remainder is covered by funding from DHBs, sponsorships, donations and fundraising from communities, profits from commercial activities, and – in the case of St John – part charges.

Current service agreements with St John and Wellington Free Ambulance (WFA) require ACC and the Ministry to pay a fixed sum per annum for three years, until 2018. This is based on the actual costs of service in 2012, adjusted each year for movements in cost indices during the previous 12 months (0.994% in 2015).

Under this arrangement St John and WFA are expected to absorb any cost impacts from demand growth and performance improvement until 30 June 2017. St John and WFA advise that their current funding arrangement does not provide them with financial sustainability and certainty for the future.

[13] Actual performance and targets differ slightly between OSJ and WFA.

Also, ambulance communications infrastructure is aging and fragile. Major investment in new technology and facilities will be needed within three years.

The funding model for emergency road ambulance services needs to be reviewed to consider that they (a) are taking on costs that would have fallen into other parts of the health system, (b) cannot sustainably absorb the increases in demand, and (c) face major investments in the near future.

Objectives

The objective of the funding review is to advise the Director General of the Ministry and Chief Executive of ACC on government funding arrangements of emergency road ambulance services that:

a) deliver a high quality emergency ambulance system, at good value for money

b) encourage effectiveness, efficiency, and innovation

c) incentivise better health outcomes and broader health system performance

d) provide greater fiscal certainty and reduced fiscal risk for government and service providers

e) provide for necessary investment in ambulance communications infrastructure

f) are affordable short-term and sustainable longer-term.

In scope

The scope of the review includes, but is not limited to, the following questions:

1. How might Ministry of Health and ACC funding for emergency road ambulance services (including emergency ambulance communications) be reconfigured to :
    - Provide greater financial certainty and sustainability for funders and providers
    - Encourage innovation and increasing value for money?
2. What relationship and accountability arrangements between funders and emergency ambulance service providers will provide for the right mix of autonomy, flexibility and oversight?
3. At what point should there be greater collaboration between St John and WFA (ie, through sharing or rationalising of functions, overheads, and infrastructure)?
4. How should non-emergency ambulance services provided to other health sector agencies and communities be funded?
5. What changes to funding and accountability arrangements in the wider health could be made to support closer working between emergency ambulance services, district health boards (DHBs), primary health organisations (PHOs), and other health service providers?
6. What is the role of commercial revenue generated from St John's and WFA's brand value in supporting emergency ambulance services?

7. Should any autonomy or funding be at risk subject to delivery of expected performance or improvements?

Out of scope

Air ambulance services.

Process and deliverables

- The review will be led by Dr Murray Horn.
- Officials from the National Ambulance Sector Office (NASO) will provide administrative support.
- A draft report will be provided to the Director General of the Ministry of Health and Chief Executive of ACC by 24 December, 2015.
- A final report will be provided to the Director General of the Ministry of Health and Chief Executive of ACC by 28 February, 2016 or other mutually agreeable date [completion date was revised to June 2016].