Making Your Impact Assessment Robust

International
Companion Animal
Management Coalition

Selecting the appropriate impacts, indicators and methods of measurement for your intervention is an important first step in conducting an impact assessment. This section further explores other factors to keep in mind as your plan and implement your impact assessment, highlighting ways of increasing the robustness of your evidence.

Ethical review

In this guidance we have selected indicators and methods of measurement that minimise animal suffering. However, it is good practice to conduct an ethical review of the indicators and methods of measurement that you plan to use for impact assessment, in addition to an ethical review of intervention itself, in particular veterinary and animal care activities that can affect welfare. Such ethical reviews are also a requirement for publication in many peer-review journals and may also be part of legislation covering research involving animals. The overarching principle of ethical review is to ensure that the potential risks are balanced by the likely outcome of the research and hence it encourages investigators to minimise harm and increase potential benefit through selection of the least invasive methods possible and good experimental design. By using non-invasive techniques such as observation (e.g. street survey and behavioural observation) this minimises the risks to the animals and the balance is favourably tipped. For a more thorough discussion of ethical reviews, and guidance on how to conduct them, see RCVS and BVA (2013).

A key question of your ethical review is whether there is the potential to cause 'pain, suffering, distress or lasting harm' through any of your methods of measurement, the threshold of pain that is used is that of introducing a hypodermic needle through the skin. In the case of taking blood samples purely for reasons of surveillance, there is potential to cause harm and an ethical review should be conducted. In some countries, such a procedure would also fall under animal testing and procedures legislation and an additional licence may be needed. Where a blood test is taken in the course of veterinary treatment and a small amount (10% or less) is used for surveillance, this may fall outside of animal testing and procedures legislation as the primary purpose was to treat the animal and no additional harm was caused. Where data is being collected from owned animals (e.g. questionnaires or clinic records) the informed consent of the owner to use the data for impact assessment should be asked and clearly recorded. This data will need to be securely stored and not shared with agencies outside of those responsible for the impact assessment, especially when the owner's details need to be maintained to allow for longitudinal study, if follow-up is not required the data can be anonymised to maintain confidentiality.

Setting up your own ethical review board is possible but may be quite time consuming. An alternative is to access an ethical review board through local research institutes, who usually have a committee already formed. As this ethical review should be conducted before the data collection starts this may also be a good opportunity to create collaboration to access further support, such as data analysis and interpretation.

Attribution and measuring your intervention effort

This guidance focuses on collecting indicator data relating to impact. However, evidence that impact has occurred needs to be presented alongside evidence of intervention effort and causality; for example, was it your intervention that caused the improvement in dog welfare or some other change in the environment? This is commonly termed **attribution**: can the change in the impact be attributed to your intervention or was this change caused by something else?

Establishing attribution can be supported from the outset by developing impacts that can realistically be achieved by your intervention, where a causal link can be logically argued between your effort and the impact (this was introduced as a ‘theory of change’ in the section on ‘Identifying Dog Population Impacts’). For example, a reduction in nuisance complaints relating to dog reproductive behaviour can be logically linked to an intervention that sterilised a significant proportion of the roaming dog population, especially if a reduction in lactating females was observed over the same time period. However, the same intervention (assuming no vaccination was involved) could not claim to have reduced rabies prevalence and any observed reduction must have been due to some other change in the environment, such as a dip in the natural oscillations of the disease. Conversely, an intervention that focused on rabies vaccination of dogs, and not sterilisation, could not claim to reduce reproductive behaviour of dogs.

Attribution can also be supported by focusing monitoring and evaluation on the intervened population only and not spreading out beyond the immediate population where impacts may become too diluted to be visible. Further testing of attribution requires robust experimental design, as described further in the following section.

Elements of robust experimental design

Although most DPM interventions are not initially envisaged as ‘experiments’, rather they are developed for social good, by utilising experimental theory they can be set-up to provide robust evidence of intervention effectiveness. This section highlights some of the key elements of experimental design that are applicable to DPM interventions. Although including all elements could be considered a ‘gold standard’ and may not be achievable for all, including any of these elements could improve the quality of any future impact assessments.

Perhaps the most fundamental element of experimental design is using a matched control. This is a population of dogs (and their associated people) where no intervention has taken place but the dogs experience a similar environment and ownership style. The dogs that have received the intervention are termed the treatment group. For example, in Jaipur the number of human rabies cases reported by the hospital from the intervened portion of the city was compared to those in the non-intervened outer ring of the city and found to be lower (Reece and Chawla, 2006); note that the intervention was spread to the entire city after a few years as the beneficial impact was so clear, hence this control no longer exists within Jaipur. An alternative is to have a control group where only a standard intervention is used and a treatment group where additional aspects are used; this is usually used where the standard treatment is expected to have a beneficial impact and withholding such a benefit from people or animals in the control group would be unethical. An example is mass vaccination for rabies control in one group of control villages and mass vaccination plus sterilisation in another group of treatment villages; then comparing indicators related to rabies control between these two groups to assess whether sterilisation contributed positively to rabies control. The approach of a standard intervention as a control and additional aspects as a treatment may be more realistic, as a complete absence of dog management is relatively rare, especially if the actions of individual owners to control their own dog's contribution to the population are defined as part of an intervention.

Note that ideally there would be several treatment groups and several control groups, termed replicates. This is because you would expect variation in all impacts over time and by having several groups of each type you will have a measure of this variation. You can then compare this to the difference between the control and treatment groups and ask the question, is this difference between the control and the treatment greater than the variation within the treatment groups? If yes, this intervention has been effective at creating the desired impacts.

Another form of control is a temporal control. This is the period before an intervention starts (sometimes called a baseline) and can be used for comparison with the period during or after the intervention. Ideally both temporal and location controls are used together, in other words the pre- and post- intervention periods are compared for both treatment and control groups as well as comparing the different responses of treatment and controls during the intervention. One example of replicated treatment and control groups combined with use of temporal controls is from Cleaveland et al. (2003); all villages (>40) in the Serengeti district were selected as treatment villages (dogs vaccinated against rabies) and 10 villages were selected randomly from the neighbouring district as control villages. The effect of vaccination in the treatment villages resulted in a reduction of dog rabies cases and suspect bites when compared with the control villages. This significant reduction in rabies cases in the treatment villages was also found when data was compared to the pre-intervention period. This finding strengthens the evidence that vaccinating dogs against rabies, lead to a reduction of dog rabies cases. The study also found that the number of dog rabies cases stayed the same in the control villages across both the pre-intervention and intervention period, and the suspect dog bites showed a slight but non-significant increase in the intervention period as compared with the pre-intervention period. The slight increase in suspect bites was due to the intervention providing an improved supply of human rabies vaccine to all villages during the intervention, hence people were more likely to seek bite treatment from their local health centre as vaccine was now available, where previously supply had been intermittent.

The most rigorous approach is the cluster randomised control trial, where the locations that will be treated/intervened and those that will act as controls are selected at random. This ensures any pre-existing differences between the locations do not bias whether they are selected for intervention. There is also replication of both controls and treatments; this is what leads to the term 'cluster'. One example of this approach is provided by Mazloumi Gavgani et al. (2002) who randomly selected a control and treatment village from 9 matched pairs of villages in the provinces of Kalaybar and Meshkin-Shahr in northwest Iran. These villages had been matched for leishmaniasis prevalence in children, so the design was a matched-cluster randomised control trial. They used deltamethrin- impregnated dog collars in the treatment villages and found a reduced incidence of leishmaniasis in dogs and children in the treatment villages compared to the control villages. By using a cluster randomised control trial the authors have removed several potential factors that could have produced this result other than the treatment itself; creating very robust evidence for the effectiveness of deltamethrin- impregnated dog collars on Leishmaniasis.

In reality, the use of a control group in dog population management is extremely rare, perhaps because the resources required to monitor and evaluate both treatment and control locations has been perceived as too large. However, we would strongly encourage this approach wherever possible as we can be more confident of attribution from the DPM intervention. Even if randomisation is not possible and only a subset of the indicators can be measured in the controls at baseline and beyond, the inclusion of a matched control can strengthen the evaluation results considerably.

As mentioned previously, the benefit of using a control group or groups is that they capture the natural variation in impacts that would have occurred over time regardless of your intervention. These variations can include confounding variables; those factors other than your intervention that influence your indicators. For example, an intervention that uses rabies vaccination of dogs to reduce human rabies cases may find that improvements in the availability of post-exposure prophylaxis (PEP) for people bitten by dogs will also reduce human rabies cases. These confounding variables should be identified when evaluating the impact of an intervention. At the planning stage, all likely confounding variables should be established so that these can be measured as well as the intervention effort. In some cases, these confounding variables can be controlled for or even avoided with good experimental design. For example, an intervention that includes providing low-cost surgical sterilisation in a deprived area may be hoping to see a reduction in the percentage of roaming lactating females. However, the percentage of lactating females may also affected by the time of year due to one or more peaks in breeding, hence time of year can be a confounding variable for some locations. Although the seasons cannot be controlled, their influence on indicators can be minimised by only comparing the percentage of lactating females recorded at the same time of year. There are many other examples of avoiding the effect of confounding variables on indicators. These include conducting street surveys at the same time of day, avoiding extremes of weather that will affect dog behaviour, introducing questionnaires in the same way and avoiding unusual days like holidays when different people will be at home.

To ensure monitoring and evaluation has the best chance of exposing a real change in an indicator, the method of measurement used needs to be reliable. The need to be reliable extends to the observers conducting the measurement as they are an important source of potential error and the Section ‘Increasing and testing observer reliability’ explains how this can be tested and minimised. Furthermore, a systematic bias could result from the desire of the observer measuring the indicators to see a change, as might be expected when that person is involved in running the intervention. They have a vested interest in wanting to see that the intervention has worked and so, even subconsciously, may record data more favourably over time. One way of avoiding this is to use independent evaluators that have no reason to want to see a change in a certain direction. Even more powerful would be to ensure the observers recording the measurements are unable to identify which is the intervention and which is a control areas – this is referred to as a blind experiment – this ensures an observer could not contribute bias to the results even if they wanted to. In reality, using blinded independent evaluators may not be possible for many interventions due to cost (although one cost effective option could be to swap staff between different interventions for monitoring and evaluation events) and also the ability to truly blind someone from an intervention when the dogs themselves may carry marks indicating they have been through the intervention. However, it is a gold standard to aim for whenever possible. Where not possible the people measuring the indicators have to be conscious of their inherent bias and fight against these, trying to remain objective throughout their monitoring and evaluation work.

Measuring intervention effort

Measuring intervention effort is essential for attributing change in impact indicators to the intervention itself, and managers need to have documented what they have done to bring about change. This is focused on the immediate results of the intervention. However, inputs, the time and resources to implement the activities, should also be measured as these will be important for assessing the efficiency of the intervention. For example, an intervention that offers low-cost sterilisation needs to record the finances required to run their clinic (the inputs), the number of dogs that come through their clinic for sterilisation and treatment (the intervention effort, see Clinic records for more details), and also what proportion of the dog population this represents in their intervention area.

The intervention also needs to have described a logical chain of steps leading from the intervention effort through to the impact (this was introduced as a 'theory of change' in the section on 'Identifying Dog Population Impacts'). By providing evidence of intervention effort and change in impact indicators, alongside a logical explanation of how this intervention has influenced the impacts, managers have a foundation from which to attribute the change to their intervention. Accurate recording of inputs will also allow additional questions about cost effectiveness to be answered.

In addition, measuring some indicators of impact will require accurate records to have been made during intervention implementation, including *when* dogs were accessed by the intervention. For example, calculating survival of unowned dogs requires knowing when the dog was last handled by the intervention (requiring a combination of individual identification and records showing when that dog received its individual mark). Ideally all this data should be stored using a database that allows for later analysis rather than on spreadsheets (which are more difficult to manipulate) or on paper. There are many database systems available, for example 'Access' available from Microsoft Office, and also online databases that can allow access from a number of computers. Whichever software is chosen the database must be backed-up to avoiding losing data.

Sampling

Sampling methods and sample size will be a consideration for all methods of measurement. This starts with identifying the target population: which dogs and/or people across what area does our intervention intend to impact upon? If this target group of dogs and associated people is small, the methods of measurement can be applied to the entire population; this would be termed a census. However, the target group is usually too large to affordably census on a regular basis and so just a sample of dogs and/or people are chosen to represent the wider group; results drawn from this sample are used to infer changes occurring in the wider group, with the acceptance that there will be a level of error in the inference as a result.

Sampling is broadly done in two ways, simple random or stratified random. In a simple random sample, every dog or person has an equal chance of being picked, for example stopping at every 10th house to ask a household questionnaire. In a stratified sample, some dogs or people have a known greater chance of being selected than others. A stratified random sample may be used when your target population clearly appears as sub-populations of different sizes with important differing characteristics related to your impact, and you don't want to run the risk of missing a sub-population when using a random selection. For example, consider a situation where dogs living in a large rural area have poorer welfare and greater disease risk than those living in a small urban area within your intervention zone. You may wish to select more dogs from the rural area for your sample and fewer from the urban area. Knowing the different chances of selection for these different sub-populations that make up your sample will be important at the analysis stage and also for repetition during future monitoring events.

Planning to apply the method of measurement to an appropriate sample size will also be important; too small and there is little chance of exposing a change in an impact indicator, too large and there will be wasted resources on unnecessary measurement. Establishing an appropriate sample size can be done statistically using power analysis, before you begin the bulk of your data collection; this essentially describes the 'power' your data will have to expose a change in the indicator, if a change has truly occurred. The power of your data will be increased by increasing the size of your sample, increasing the size of the change you want to expose (e.g. a reduction in the percentage of emaciated or thin dogs in the population by 20%, rather than 10%) and increasing how risky you are willing to be about the accuracy of results (e.g. will you accept being 80% confident that the results are accurate? In other words, is 20% an acceptable risk that your results are actually wrong?). Conducting power analyses is usually done mathematically and will require the support of a statistician who should also be able to advise you on what statistical tests will be most suitable and most powerful for your data. They will need to know the indicator you are interested in and how it will be measured. What is the size of the effect that you want to measure, so what is the baseline value of your indicator and what target do you have in mind as a definition of success, e.g. a 10% reduction from baseline of 50%? How much risk are you willing to accept regarding the accuracy of the results? They will also need some idea of variability or error in your measurement; for example, if you repeatedly measure the body condition score of the same group of dogs over a very short time frame, how much does the % of the population scored as thin or emaciated change? This last question can be answered by pilot testing your methods on dogs or people. Note that improving your measurement to reduce any errors will also help increase the power of your data to expose a change.

Preparing and conducting power analysis clearly requires time and the resources of a statistician, but is the ideal approach to ensure your monitoring has the best chance of resulting in accurate evaluation. Where this is not possible a general approach is to increase the sample size to the maximum you can afford whilst still being able to repeat your measurements over time, which is the core characteristic of evaluation.

Consistency in method

Selecting an appropriate size and composition of sample will help to ensure the data resulting from your measurement is accurate and representative of your target group of dogs and/or people. However, perhaps even more important is maintaining a consistent and precise protocol for how the method of measurement will be conducted, as this will reduce error in your data resulting from variations in how the method was carried out. Even where a sample is accidentally biased towards a group of dogs or people, if the selection criteria and method are consistent, the data will accurately expose a change in this biased group, if such a change has truly occurred.

Inconsistency can come from many sources, including differences between observers (see later section on 'Increasing and testing observer reliability' for an example of this) and failure to develop and consistently apply a standard method, leading to changes in the sample selected or protocol used for measurement over time (e.g. the introduction used by an interviewer when conducting a questionnaire changes over time because it was not written down initially, a different introduction can influence the answers given by the interviewee).

Ensuring there are complete and accurate records of all methods used, and allowing time to familiarize and train people to the same standard, will help to reduce error in the data. Logistically this may be supported by assigning a lead person(s) to be responsible for monitoring and evaluation of the intervention with consistency as a key indicator of their performance.

Increasing and testing observer reliability

The aim of monitoring and evaluation is to explore and expose changes in targeted impacts. In order to do this, methods need to measure changes in the indicators as precisely as possible. One of the challenges to this will be reliability in measurement (first introduced in the section ‘Attribution and measuring your intervention effort’). A potential source of error that can be mitigated will be the effect of the person that conducts the method of measurement. As far as possible, the staff involved should be consistent and make every effort to score in the same way at each monitoring event. However, staff changes are inevitable and people can also unintentionally drift in their approach and assessments, especially as some changes will be slow and will require several years of monitoring to become apparent. In order to overcome this challenge, every monitoring event (even if staff are the same) should include a period of refresher training where the protocols are reviewed and discussed in detail. Ideally, a bank of photos is built up of dogs in varying body and skin condition that can be used for refresher training on scoring the body and skin condition of dogs.

Further, key indicators can be checked for inter-observer and intra-observer reliability. Inter-observer reliability is a measure of how much agreement, and disagreement, there is between the different people involved in conducting the method of measurement. Intra-observer reliability is a measure of how consistently the same person scores over time. The following is an example of testing inter-observer and then intra-observer reliability in scoring body condition. This should be carried out before any monitoring event and at regular intervals:

Body condition score training and agreement test

The body condition score recommended in this guide is a 5-point scale (Annex A) requiring observation only with no need for physical examination. The indicator produced through body condition scoring is the percentage of the adult (non-lactating) population that is body condition score 1 (emaciated) or 1 and 2 (emaciated and thin). Methods of measuring body condition include street surveys and clinic records.

Before either method of measurement is conducted, all observers that will be scoring dogs should review the protocol and category descriptions in the body condition score tool (Annex A), and together discuss and score at least 20 dogs of varying condition to ensure they have the same understanding of the tool. This can either be done using photos of dogs or by directly observing dogs in the street, shelter or clinic environment. A powerpoint presentation to help with this training titled 'Dog body condition scoring using visual assessment' is available on the www.icam-coalition.org website. Ideally at least some direct observation of live dogs should be done in the environment matching the method of measurement they will be using; i.e. in the street if they are to score dogs during a street survey and in a clinic if they will be scoring as dogs pass through an intervention. This should be done both by new observers, and people who have conducted these methods before, in a group together.

Once they have taken part in the training and agreed through open discussion, the body condition scores for at least 20 dogs, observers can then take part in an inter-observer test. The following is a suggestion for how this test could be run, based on the process developed by AssureWel (www.assurewel.org):

Inter-observer test

The observers are asked to score test sets of 10 photos of dogs. You can develop your own test sets or you can access test sets in the form of an online quiz at www.icam-coalition.org. This online quiz is comprised of dogs falling into different categories of body condition score. The photos are presented in a random order and the observers select the body condition score for each dog. The quiz gives you immediate feedback on whether your assessment was correct or incorrect.

If the observer scored less than 9/10 correct, they are advised to review the powerpoint presentation 'Dog body condition scoring using visual assessment'. You can also discuss the photos in this presentation to help identify anatomical features that may have been missed. They can then retake the quiz. If they scored 9/10 or 10/10 they also need to retake the quiz again as observers need to score 9/10 or 10/10 on two consecutive sets to pass the test and be considered proficient at visually assessing body condition score. If they score below the 9/10 threshold on a set they need to start again with at least a further 2 sets before 2 consecutive pass scores can be achieved. For an observer who scores 9/10, or more, on two consecutive sets you can be 85% certain that they can score dogs for body condition with at least 80% accuracy (binomial distribution, with *n*=10 and *p*=0.80).

As the aim of monitoring is to evaluate change over time, consistency in scoring over time (intra-observer reliability) is just as important as agreement between observers (int¬er-observer reliability). Hence observers should retake the quiz at the outset of the next monitoring event following a period of refresher training.

Test sets can also be built-up specific to a location. Ensure that photos are at high resolution and in sharp focus (this is especially relevant for scoring the presence of ribs) so that they can be shown on a large computer screen or projected. The full body of the dog should be visible, showing both one side and at least the lower part of the back, so that both hip bones and the vertebrate can be seen if prominent, plus the extent of the waist. Ensure that dogs from the full range of categories are present in the test set in approximately the proportion you would expect to see in the location.

Once in the field, reliability can be informally tested by observers working pairs and asking each other for confirmation of a score. Discrepancies can be talked through and where agreement can't be reached the observers can refer back to the original definitions of each score and even take a photo of the dog(s) for discussion with a wider team after observation is concluded.

Using your results

This guidance aims to help intervention managers to decide *how* to measure their impact, by selecting the most meaningful indicators and suitable/affordable methods. However, perhaps the hardest job falls to the intervention managers to ensure monitoring is actually done; that time is made for analysis and interpretation; learning and improvement occurs, along with dissemination of results to others.

This process can be helped by developing a monitoring and evaluation plan. This can include: a detailed description of each indicator along with the impact it measures; a detailed protocol for the method of measurement and associated budget and timeline for when this should be conducted; name(s) of people that will be responsible for ensuring the method is conducted and data reported; and finally a plan for regular but infrequent evaluation 'events'. Evaluation events are workshops where relevant project staff and representatives from wider stakeholders, potentially including donors, come together to review indicator data and assess the extent, or barriers, to change over time; resulting in an impact assessment of the intervention and suggestions for improvements.

Commitment to monitoring and evaluation will also be greater if designed from the outset as an opportunity to learn as opposed to the need to prove impact to external audiences. This uses as a starting point the learning that intervention field workers themselves need in order to implement the intervention more effectively, rather than the results that managers need to demonstrate impact to senior or external people. The concept is that evidence of impact will be an emergent property of the learning, rather than the other way around and is termed 'learning-based monitoring and evaluation' rather than 'results-based monitoring and evaluation'.

Part of the monitoring and evaluation process includes a phase of analysis and interpretation, requiring the support of someone with an understanding of data analysis. We recommend, if such expertise does not exist within the intervention team, that external scientific expertise is sought, potentially from universities, research institutes or donors, before monitoring begins. Building such a relationship from the outset ensures that data is collected in a way that supports later analysis to answer questions about how indicators have changed. For example, using sufficient sample sizes and using protocols that minimise potential confounding variables or at least concurrently collecting data on these variables so that their effect can be tested.

Returning to the subject of evaluation events, these set a deadline for when all the relevant data should be available for interpretation and learning. This helps to ensure those people collecting indicator data can see that it is being valued. It also ensures that data is not simply collected for many years, but that it is analysed and actually used for learning on a regular basis. The evaluation event can also finish with a reporting phase, to all intervention staff, community and government representatives, and donors. Wider dissemination to other interventions via project evaluation reports, media releases, conference presentations and peer-reviewed publications would allow more widespread learning. The ICAM Coalition in particular would value receiving such project evaluation reports and any information on performance of indicators/methods of measurement mentioned in this guidance or innovated by the intervention team.