Working paper 3

Learning lessons from challenges encountered in evidence-based practice in other sectors

Project	Standards of evidence in housing with care, support and health
Identifier	Working paper 3
Title	Learning lessons from challenges encountered in evidence-based practice in other sectors
Author	Jim Vine, Director of Evidence, Data and Insight, HACT
This revision	1.0
Date	2015-12-01
Revision history	0.9: working version for project partners. 2015-03-20 0.10: (internal draft) correction of minor errors; other small edits for clarity. 2015-11-30 0.11: (internal draft) added map; changed reference to validity per Lilienfeld et al 2015; added some references. 2015-12-01

Contents

1. Introduction

Other sectors have far more experience of evidence-based practice than housing. It make sense to learn from their successes and the techniques that have been developed and deployed to understand which types of evidence are best for answering which questions. But even in those fields that are most advanced in their use of evidence there are problems, and it makes sense to learn from those too. As housing takes some early steps on a journey towards a more evidence-based future it should not slavishly follow the path that others travelled, but instead learn from their missteps.

Medicine is probably the most advanced evidence-based field, and yet even there, the situation is not perfect, as described by Ben Goldacre:

> *"Drugs are tested by the people who manufacture them, in poorly designed trials, on hopelessly small numbers of weird, unrepresentative patients, and analysed using techniques that are flawed by design, in such a way that they exaggerate the benefits of treatments. Unsurprisingly, these trials tend to produce results that favour the manufacturer. When trials throw up results that companies don't like, they are perfectly entitled to hide them from doctors and patients, so we only ever see a distorted picture of any drug's true effects."*[1]

The incentives in social housing will be very different from those affecting drug development, and the absence of the high stakes profit motives that relate to potential blockbuster drugs should reduce the incentives towards deliberately flawed evidence creation. But many other potential problems will apply even in the most altruistically-driven parts of the sector, and those parts of housing providers' business where competition between them is virtually non-existent.

The problems encountered in existing evidence systems described in this paper should not be read as a counsel of despair. In spite of the flaws in its systems, medicine has continued to improve and has delivered significant health benefits to millions. Even a flawed evidence ecosystem can provide many beneficial impacts. Rather they are presented to provide the opportunity for housing to enjoy the benefits of following other sectors, allowing it to establish its own systems so they avoid these problems from the outset, not replicate them and have to resolve them in decades to come.

Furthermore, because other sectors have been conscious of some of these flaws for many years, there are also many solutions that have been proposed that might form the basis for thinking in the housing sector.

[1] Goldacre, B., 2012. Bad Pharma: How Medicine is Broken and How We Can Fix It. (In 'Intro'). Fourth Estate. London.

2. The problems

Problems identified as having been encountered in other sectors adopting an evidence-based approach include:

Publication bias

Publication bias refers to the tendency for studies that find a positive association between an intervention and an impact to be more likely to get published than ones that find no effect.

There is always a potential for 'false positive' results in evaluation (i.e. trials that indicate something is effective when in reality it is not). At the extreme, it could be possible for there to be 20 trials conducted of some intervention, and if the 19 showing it to be ineffective sit unpublished but the 1 that shows it is effective gets published, an entirely inaccurate impression of the state of the intervention could emerge. Perhaps more commonly, early trials going unpublished might result in later ones being conducted needlessly, when earlier publication of the non-effective findings could have freed up subsequent researchers to focus on novel interventions rather than duplicating work without knowing it.

In the medical field, where trials are typically conducted for drugs under the control of a particular company, there might be motivations for that company to leave unflattering results unpublished, and deliberately skew towards the publication of positive results. In sectors like social housing, it is far less likely that there will be large financial incentives for particular interventions to appear positive, especially in those parts of housing providers' businesses that do not directly compete with each other, but other factors that can lead to publication bias could still apply. Positive findings may simply appear more interesting, important or publication-worthy,[2] and there may be reputational benefits (if not financial ones) associated with being attached to a positive novel intervention, and these potentially biasing factors may apply across sectors, irrespective of the relatively low influence of profit motives.

Shifting primary indicators

The primary indicator is the main outcome measure that you are examining to see whether your intervention results in an improvement. If data are collected on a large number of indicators and the primary one is selected after the trial there is potential for "cherry-picking", where the authors select the indicator demonstrating the most positive results. The probability of false positive results increases as the number of potential indicators increases: the more things you are measuring the greater the chance that one of them will improve for the intervention group compared to the control group by pure chance.

[2] Peplow, M., 2014. Social sciences suffer from severe publication bias. Nature News & Comment. http://www.nature.com/news/social-sciences-suffer-from-severe-publication-bias-1.15787

Practitioner difficulty in establishing the state of evidence

There can be multiple layers that act to make it difficult for practitioners to establish the current state of evidence on a given topic.

Assuming a particular piece of evidence has been published somewhere (see the section on publication bias), it may still be somewhere that is simply inaccessible to practitioners. One particular issue of this kind is evidence published in many peer-reviewed academic journals; for those without access to these journals via a university subscription, access to individual articles can often be prohibitively expensive.

Even where practitioners and decision-makers could gain access to a particular paper or report, there can still be challenges related to actually doing so, and in terms of meaningfully engaging with the content. One issue could be the technical skillsets necessary to find papers (which might be located in academic databases that will be unfamiliar to many practitioners) and to read and interpret them, if they are presented in unfamiliar technical formats. Furthermore, even those practitioners that have acquired the necessary skills may find that it is highly time-consuming exercising those skills to thorough research the state of the existing literature on a given topic.

Sometimes the issue might simply be that the study is poorly reported. Evidence has been found across a range of fields that reporting of trials can be suboptimal.[3]

Non-generalisable findings

Other fields, including medicine, contain examples of studies being conducted to very high standards, and producing findings that are highly robust in their own right (described as having high internal validity), but that cannot be generalised to the situation that pertains in real practice.

The factors that lead to this poor generalisability can include the inclusion criteria for people included in the trial being too tightly drawn, meaning that the findings only relate to a very small subset of the population, whilst practitioners are interested in how the intervention performs for a much wider group. Some studies of this issue have found that criteria disproportionately exclude those from vulnerable backgrounds, raising ethical concerns.[4]

Another way in which a study might have high internal validity but low external validity (i.e. low generalisability) is if the intervention is implemented in a very tightly controlled and highly resourced

[3] Rios, L.P., Odueyungbo, A., Moitri, M.O., Rahman, M.O., Thabane, L., 2008. Quality of Reporting of Randomized Controlled Trials in General Endocrinology Literature. The Journal of Clinical Endocrinology & Metabolism 93, 3810–3816. doi:10.1210/jc.2008-0817 http://press.endocrine.org/doi/abs/10.1210/jc.2008-0817

[4] Trivedi, R.B., Humphreys, K., 2015. Participant exclusion criteria in treatment research on neurological disorders: Are unrepresentative study samples problematic? Contemporary Clinical Trials 44, 20–25. doi:10.1016/j.cct.2015.07.009

way in the trial – perhaps with extensive support from the research team – that would be impractical in any roll-out at scale.

It should be noted that non-generalisable findings are not always problematic: they may simply reflect the fact that the context in which an intervention is delivered is important, so it genuinely does work in some places and not in others.[5]

Mismatch between practical significance and statistical significance

The amount of impact from an intervention that will be deemed to be statistically significant is dependant, generally speaking, on the sample size (the number of people involved in the trial); a large sample size will be able to detect small impacts with statistical significance whereas a small sample size will mean that the impact is only statistically significant if it is large.

If there is a mismatch between the choice of detectable effect size that is the basis for determining statistical significance and the size of effect that would be practically significant it can result in the trial being overpowered or underpowered.

An overpowered trial is one that is capable of detecting effect sizes that are too small to be of practical significance; for example, a trial might be capable of detecting an effect of an intervention of a 0.1% impact on some outcome, when in practice no one would deploy the intervention if it was going to deliver less than, say, a 10% impact. Overpowered trials are wasteful of resources (since it invariably costs more to conduct a trial with a larger sample size) and involve more participants in research than is necessary.

Underpowered trial create the opposite risk: not being able to detect as statistically significant an effect that would be meaningful in practice; for example, a practitioner might be willing to consider deploying the intervention if it offered a 10% impact, but an underpowered trial might only be capable of detecting an impact if it is above 30%. These trials are also potentially wasteful, as they can produce inconclusive results, and the finding that the impact was statistically insignificant might be taken to mean that it was negligible.

At the extreme, it has been suggested that overpowered studies can crowd out a multiplicity of research that might have more practical significance.[6] For some medical conditions there are a limited number of patients who might be available to participate in a trial, but an analogous argument might be advanced where the constraint is the budget available to recruit participants and deliver altered services to them: if the budget will stretch to delivering a novel intervention to 500 people, is it better to have one treatment arm with 500 people or five treatment arms with 100 people each? Naturally there may also be operational issues that could influence the number of

[5] See, for example: White, H., 2015. Learning what works and why from evidence of what doesn't work. http://www.alliance4usefulevidence.org/learning-what-works-and-why-from-evidence-of-what-doesnt-work/

[6] Horrobin, D.F., 2003. Are large clinical trials in rapidly lethal diseases usually unethical? The Lancet 361, 695–697. http://solvingkidscancer.org/fileread?t=facts_articles&i=13

different interventions that it might be practicable to test at once, but understanding the size of effect that would matter should play a part in designing studies.

Use of surrogate outcomes

Surrogate outcomes are used when the outcome of interest is hard to measure – perhaps that it will take a long time to become visible. For example, in a health study, if the outcome of interest is lowered incidence of heart attacks, but the target population is expected to have an increased rate of heart attacks that would be detectable over a period of many years rather than months, the study might choose instead to look at some other marker that is thought to correlate with heart attacks.

The potential problem of surrogate outcomes occurs if these markers turn out not to be such good indicators of the actual intended outcome as was previously believed. In one notable example in medicine, anti-arrhythmic drugs (drugs designed to stabilise an irregular heartbeat) were found to be effective at the surrogate outcome of stabilising the heartbeat; it was thought that because an irregular heartbeat is normally associated with a higher risk of heart attack and death that this stabilisation would lower death rates. However, when longer-term studies looked at actual measures of interest (heart attacks or deaths) they found that patients treated with the anti-arrhythmics actually had a significantly elevated risk of death.[7]

Lack of comparison of related interventions

In medicine it is common for there to be a proliferation of drugs designed to treat a particular condition using similar mechanisms. These so-called 'me-too' drugs may be developed not because of one drug company copying another but because all the variants were in development at the same time and it was a matter of timing which arrived to market first and which became the 'me-too'. Due to the incentive structures for drug companies, and the fact that drugs tend not to compete on cost effectiveness, these variant drugs are rarely tested against one another so there is often no information on comparative effectiveness.

Novel and un-validated outcome measures

Inappropriate selection of the outcome measures being recorded can result in generating evidence that is hard for other sectors to use. Novel outcome measures that are not subject to appropriate validation can also present risks associated with their designs not avoiding risks such as those related to self-reporting.

[7] Epstein A E, Hallstrom AP, Rogers W J, Liebson P R, Seals A A, Anderson J L, Cohen J D, Capone R J, Wyse D G. 1993. Mortality Following Ventricular Arrhythmia Suppression by Encainide, Flecainide, and Moricizine After Myocardial Infarction: The Original Design Concept of the Cardiac Arrhythmia Suppression Trial (CAST). JAMA. http://jama.jamanetwork.com/article.aspx?articleid=409358

Inability to identify why an intervention does not work

An evaluation of an intervention that solely examines the intended outcome of interest using an RCT methodology or similar will only be able to provide evidence of the extent to which it works or not; it will not provide information on why it did not work if it turns out to be ineffective.

This limited ability to identify the reasons for underperformance of interventions that prove ineffective when evaluated constrains those designing programmes from learning lessons, where tweaking might increase the effectiveness of an approach.

Challenges comparing interventions due to variety of potential outcomes

Many forms of decision-making that will seek to draw on evidence will amount to the allocation of scarce resources. If making a decision between two interventions targeting the same outcome, evidence of their relative costs and relative effectiveness at achieving the outcome should be quite directly applicable to the question at hand. For example, if two programmes are aiming to get tenants into work, and you know that one is likely to have a 50% success rate for the people you are hoping to deliver it to at a cost of £2000 per participant and the other has a 20% success rate for a cost of £3000 per participant, it is easy to see which is most cost effective at delivering job outcomes. However, housing providers undertake a wide variety of activities, and decisions ultimately need to be made not just between different projects targeting the same outcome, but between whole areas of activity that might be aiming to deliver quite distinct outcomes.

Difficulties in successfully scaling pilots

In the International Development sector, examples have been identified of projects that performed well when evaluated in pilot studies but significantly underperformed when scaled up. The potential causes of this have been identified as:

- Weaker implementation at scale;
- Weak fidelity to programme design (where the intervention is not carried out at scale exactly as defined in the specification);
- The pilot was in a specific context;
- The programme has already reached all those who can benefit.[8]

Evidence over-ridden by other concerns

Even in a context where evidence is produced and accessible, decisions may still be taken that do not accord with the current evidence, driven by political concerns. In the NHS a special Cancer Drugs Fund was established outside of the normal evidence driven approaches. It has been estimated that the £230 million spent on this Fund in 2013/14 delivered modest survival benefits (commonly of two or three months) to 19,282 patients, amounting to less than 3,400 quality-adjusted life years

[8] White, H. (2015) What's wrong with evidence-informed development? Part 2 http://blogs.3ieimpact.org/whats-wrong-with-evidence-informed-development-part-2-2/

(QALYs); invested in the general NHS budget, the same money could have delivered over 17,800 QALYs.[9]

[9] Claxton, K. (2015) The UK's Cancer Drugs Fund does more harm than good http://www.newscientist.com/article/dn26785-the-uks-cancer-drugs-fund-does-more-harm-than-good.html

3. And some (proposed) solutions:

Evaluation registries with prospective registration

> *"The purpose of clinical trial registration is to prevent selective publication and selective reporting of research outcomes, to prevent unnecessary duplication of research effort, to help patients and the public know what trials are planned or ongoing into which they might want to enroll, and to help give ethics review boards considering approval of new studies a view of similar work and data relevant to the research they are considering."* [10]

Clinical trials registries have been created with an intention of improving transparency about the trials that are being conducted and their results. They make it easier to access information on trials, by holding it in a standardised format[11], by being accessible to the public at no charge, and by being electronically searchable.

Through prospective registration, where trials should be registered before they start, registries are intended to help to address the issue of publication bias. If all trials were registered before they commenced it would be harder for those with vested interests in showing an intervention to be effective (such as drug companies testing new treatment) to get away with running multiple trials and selectively publishing the results of only those with positive results. It would also help to mitigate the impact of the other sources of publication bias, as future researchers would be able to see that a trial had already been conducted on a subject, and even if the results have not been published they could enquire to the trial team about its results.

By requiring information to be registered on details such as the primary indicator, registries can also help to address the issue of shifting primary indicators.

For the full potential benefits of registries to be acquired it would be necessary for all studies to be registered in advance. In medicine this is promoted by a policy of the most prestigious medical journals (ICMJE) requiring that registration of clinical trials in a public trials registry at or before the time of first patient enrolment as a condition of consideration for publication. The ICMJE also defines that good registers should be accessible to the public at no charge, open to all prospective registrants, managed by a not-for-profit organisation, have a mechanism to ensure the validity of the registration data, and be electronically searchable.[12]

---

[10] International Committee of Medical Journal Editors (n.d.) Clinical Trial Registration. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

[11] For example, the WHO Trial Registration Data Set, http://www.who.int/ictrp/network/trds/en/

[12] International Committee of Medical Journal Editors (n.d.) Clinical Trial Registration. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

Prospective registration of trials is also mandated by some funders of medical research.

It should be noted that although steps are in place in medicine to encourage prospective registration, progress is still needed to achieve full compliance with the requirements, especially around timeliness of registration.[13]

It should also be noted that in the creation of evidence related to housing, publication in a prestigious journal may not be a particular goal, and the mechanisms employed in health research may not be available to encourage registration in this sector. Further consideration will be needed to ensure the creation of a culture that values the accessibility of research; particular attention might be warranted around creating incentives that encourage the publishing of null findings.

Open Access publishing of findings

Academic publishing has proved to be a lucrative business for many years: journal publishers generally rely on a lot of volunteer labour provided by academics in the form of editorial boards and peer reviewers, yet charge expensive subscription fees for journals and access fees for individual articles for those not attached to a university with a subscription.

An alternative model has emerged in recent years, known as Open Access, whereby the costs are borne by those submitting articles, with the outputs being freely accessible to all-comers. Many publishers now offer options for Open Access, even within journals that are otherwise closed; one such scheme, Sage Choice, for example, levies £1,600 'article processing charges' (APCs) in the Science, Technology and Medical fields, with APCs of £800 in the Humanities and Social Sciences.[14]

Evidence curation, summarisation and review

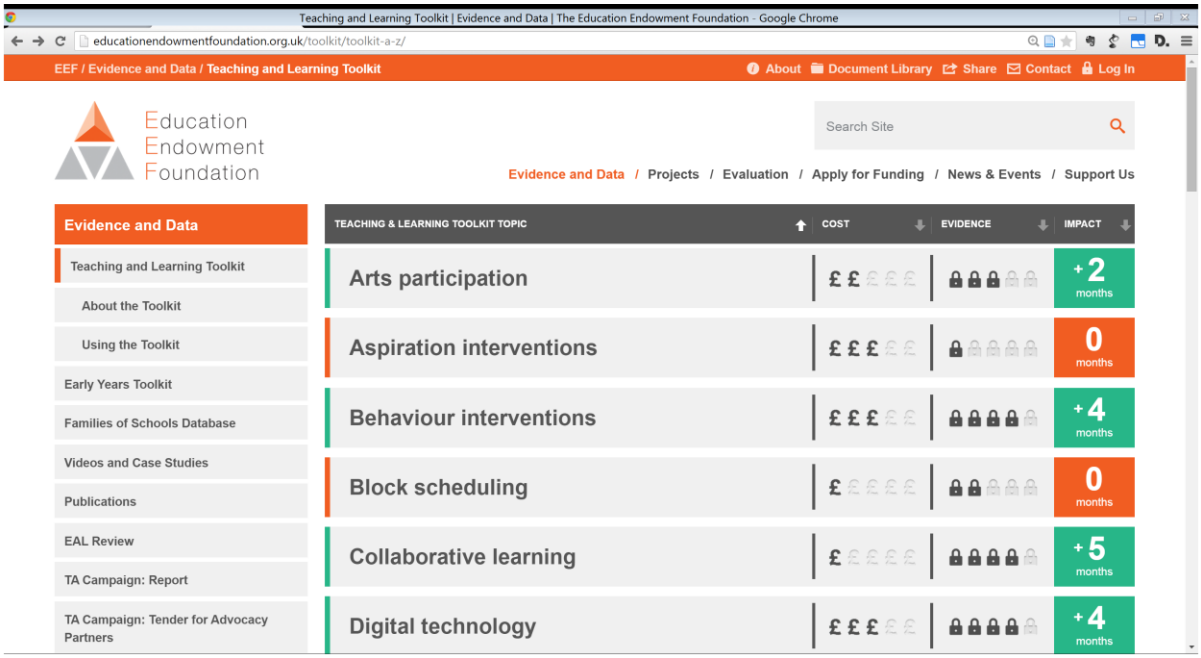
A number of different approaches to collecting evidence and making it accessible in summary format have emerged over the years. Examples of different models include those that provide broad overviews of a range of interventions and those that provide specific guidance on a particular issue.

The Sutton Trust-EEF Teaching and Learning Toolkit is an example of an overview that summarises the current state of evidence, covering both the level of effectiveness indicated for a particular intervention and an assessment of the robustness of the evidence.[15]

[13] Huser, V., Cimino, J.J., 2013. Evaluating adherence to the International Committee of Medical Journal Editors' policy of mandatory, timely clinical trial registration. Journal of the American Medical Informatics Association 20, e169–e174. doi:10.1136/amiajnl-2012-001501 http://jamia.oxfordjournals.org/content/20/e1/e169

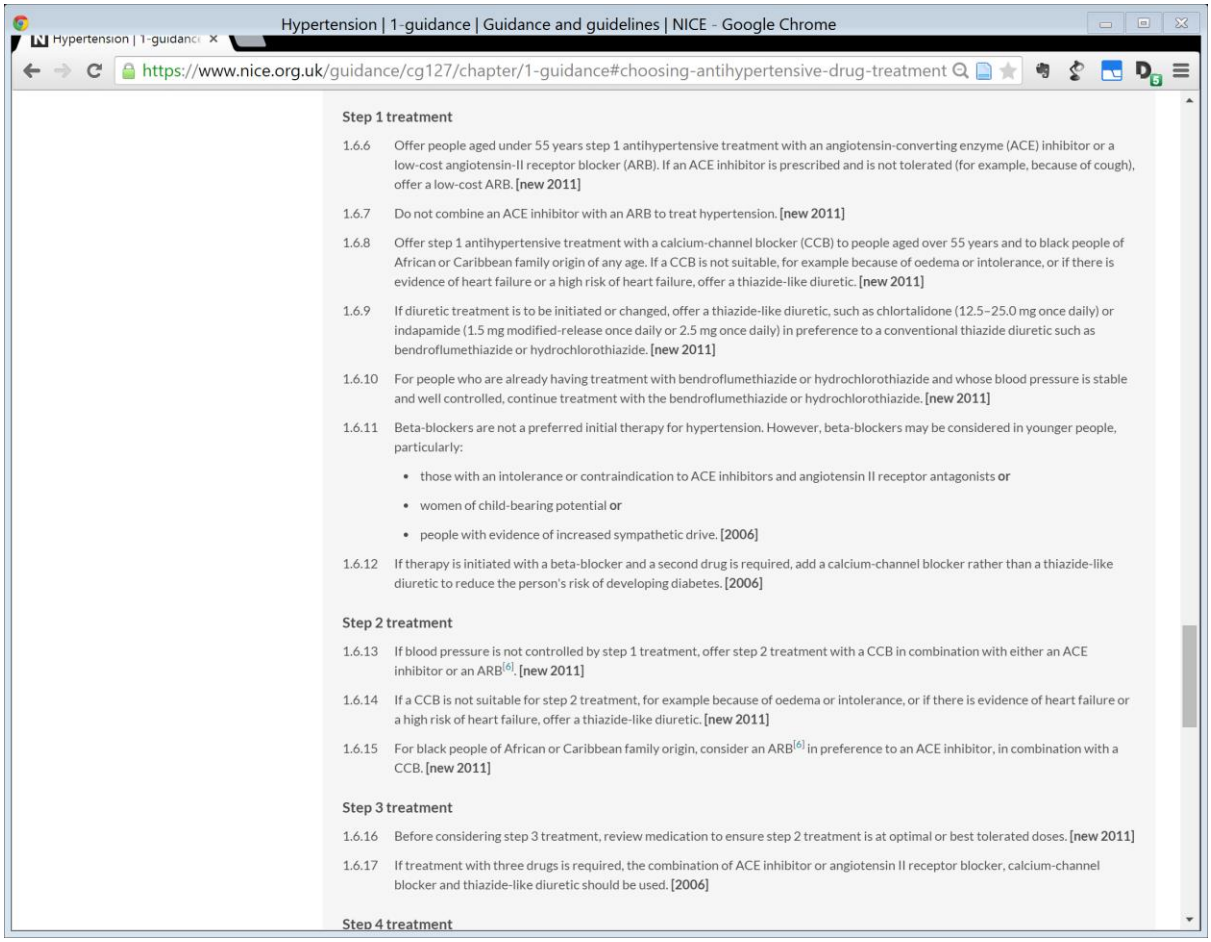
[14] http://www.uk.sagepub.com/sagechoice.sp

[15] https://educationendowmentfoundation.org.uk/toolkit/toolkit-a-z/



An example of a form of summary that specifies the particular intervention(s) to adopt can be found in NICE Guidance, which includes sections that indicate the treatments to use for a condition. In some areas this may highlight a series of options to be tried, ranking them as 'first line', 'second line', etc., and may include different recommendations for different patient groups. See, for example, section 1.6 of Clinical Guideline 127 on the clinical management of primary hypertension in adults.[16]

[16] See, for example, section 1.6 of NICE (https://www.nice.org.uk/guidance/cg127/chapter/1-guidance#choosing-antihypertensive-drug-treatment-2)



Evaluations designed in a pragmatic mode

A distinction has been drawn between trials conducted with pragmatic or explanatory attitudes.[17] Briefly, an explanatory attitude implies the researcher is trying to identify some fundamental scientific fact, or the presence of effects under idealised conditions, whereas a pragmatic attitude is adopted by a researcher trying to understand what will happen in real practice. Both forms of research are valid to advance the state of knowledge, but in terms of their ability to contribute meaningfully to practice decisions, pragmatic trials are likely to be more useful for practitioners.

Pragmatic evaluations will seek across the aspects of the trial to maximise their applicability to normal practice. The key aspects will typically comprise the population, intervention, comparator, outcomes and the target (sometimes referred to by the PICOT acronym).

- The **population** that the trial is conducted in should be as close as possible to the population that practitioners will ultimately want to deploy the intervention in, so if the intervention is going to be aimed at older tenants in general needs housing the trial should probably not be

[17] (Reprint of a 1967 article) Schwartz, D., Lellouch, J., 2009. Explanatory and Pragmatic Attitudes in Therapeutical Trials. Journal of Clinical Epidemiology 62, 499–505. doi:10.1016/j.jclinepi.2009.01.012 https://www.med.upenn.edu/sleepctr/documents/ExplanatoryandPragmaticAttitudesinTherapeuticTrials_JClinEpidemiol_2009.pdf

conducted in a sheltered housing scheme, even though that might be a more convenient way of securing a population in the broad target age band.

- Delivery of the **intervention** in the trial should be similar to eventual practice. This would imply, for example, that those delivering a service in the trial should not get substantial additional training in how to deliver it or hands-on support from the research team, if those forms of support would not be available for any subsequent roll-out.
- The **comparison** made in the trial should be against a real alternative practice. Whereas an explanatory trial would be interested in testing an intervention against 'no service', to detect whether the intervention has any effect, a pragmatic one would be more interested in testing against current practice to see whether the new service out-performs business-as-usual.
- The **outcomes** of eventual interest, either to the housing provider as an organisation or the service user, should be tested in a pragmatic trial, as opposed to an explanatory trial might be satisfied with learning about the impact on some intermediate outcome. If a community investment project on CV writing is intended to support tenants into work, a more pragmatic trial would seek to assess whether people got jobs, whereas an explanatory one might simply assess whether those completing the project had improved knowledge of CV preparation.
- Whilst the most obvious **target** for a study might be to detect whether a new intervention is better than the existing service (superiority), in some circumstances it might be appropriate to test for equivalence or non-inferiority; this might be the case where the proposed new intervention is better in some other respect (perhaps cheaper or easier to deploy) and so it only needs to be demonstrated that it is not worse than the previous option for it to be rolled out.

Causal chain analysis and theory-based impact evaluation

In the International Development sector, 3ie (International Initiative for Impact Evaluation)[18] promotes an approach to evaluation that it describes as theory-based. By considering a causal chain, the evaluator is able to think about all of the points at which the link between the intervention and the primary outcome might break down. If it is possible to collect secondary outcome data around some of the links in the chain, it might be possible to identify 'why not', if an intervention fails to achieve its desired impact.

In the case of a CV-support workshop to support tenants into sustained employment, a simple sketch of some of the links in the chain might be:

[18] http://www.3ieimpact.org/

1. Enrols in project offering CV support
2. Attends CV workshop
3. Increases knowledge of how to construct a good CV
4. Constructs a good CV
5. Sends new CV off to relevant potential employers
6. Gains interviews
7. Gains job offer
8. Accepts job
9. Retains job

If secondary outcome data were collected on all of those stages and it was found that those on the programme were not experiencing improved employment outcomes, the secondary outcomes could be examined to gain insight into the problem. It might, for example, highlight that the participants are gaining more interviews than they otherwise would have, but that these interviews are not translating into job offers. This might provide information to suggest that it could be worth testing a new version of the project that offers both CV and interview support. Similar potential amendments can be identified for any point on the causal chain where the performance is found to drop off.

Collection of secondary outcome data for this purpose is legitimate and often desirable, and should not be mistaken for collection of a range of outcomes data simply to have more potential chances of being able to declare an intervention a success on one of many measures.

As identified in Working Paper 1,[19] the use of appropriate robust qualitative research methods can also be very valuable in seeking to understand more fully the 'how' and the 'why' (or 'why not') alongside the quantitative measures identifying whether an intervention works.

Involvement of practitioner staff expertise in considering outcome options

In medical research there is a concept of minimum clinically important difference (MCID). Below the MCID, any improvement would be considered negligible and not worth acting upon. Engagement with practitioners and others to identify what level of improvement would be considered a tangible benefit is a sensible contribution to avoiding overpowered studies that are set up to capture tiny differences that practitioners would not be interested in.

Such an approach would also help to ensure that resources are appropriately allocated to identifying changes that matter: for a given budget to test potential interventions it would normally be preferable to run several small trials that can only detect meaningful practical differences in outcomes rather than one large one that could detect even a tiny impact.

[19] Vine, J., 2015. Introduction to research methods and considerations of their relative strengths. http://www.hact.org.uk/sites/default/files/Intro%20to%20research%20methods%20-%20Stds%20of%20Evidence%20WP1%20v1.1.pdf

Consistent thorough reporting of evaluations

Making study findings accessible naturally begins with ensuring that the results are published in some form; this might be supported through agreements between researchers and commissioners that results will be published on completion, irrespective of the findings. Beyond ensuring findings are reported in the first place, increased consistency of reporting has been proposed to ensure that necessary elements are included in publications to permit them to be fully usable.

Frameworks such as the CONSORT Statement have been developed to provide "a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation."[20] Similarly, the PICOT acronym (population, intervention, comparison, outcomes, target) provides a set of headings that could usefully be reported consistently for every study to ensure that readers have sufficient information on the relevant aspects of what was being tested. It has been shown that framing of the research question using the PICOT format is positively associated with better overall reporting quality.[21]

A clear description of the intervention being evaluated that adequately communicates exactly what was done, by whom, to whom and in what circumstances, ideally with a cost breakdown (or information that would enable that to be estimated) will greatly enhance the likelihood that potential future users of the evidence will be able to make use of it.

Reporting should also be designed such that it facilitates the answering of the 'so what?' question, i.e., to allow potential users of the evidence to easily identify what the implications are for practice or commissioning.

Evaluate at scale as well as in the pilot phase

Because of the potential for differences in context between an initial small pilot and a subsequent roll-out, it may be necessary to conduct further evaluation of projects as they are scaled up.[22] The more pragmatic the mode of the study in the initial stage, the more likely its results will be generalisable at scale, potentially avoiding the need for repeated studies. However, there might be operational reasons necessitating a more explanatory approach at the pilot stage; in this case the pilot could be considered an efficacy study (i.e. seeking to identify a detectable effect under particular conditions), generating useful knowledge that may indicate an intervention merits further development and testing. It would not, however, negate the need for a study that tests effectiveness in conditions closer to real practice.

Other approaches can also reasonably be adopted for evaluations at different stages of an intervention's development. Early stage pilots may simply seek to test the feasibility of delivery and

[20] http://www.consort-statement.org/

[21] Rios, L.P., Ye, C., Thabane, L., 2010. Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials. BMC Medical Research Methodology 10, 11. doi:10.1186/1471-2288-10-11 http://www.biomedcentral.com/1471-2288/10/11

[22] http://blogs.3ieimpact.org/whats-wrong-with-evidence-informed-development-part-2-2/

its acceptability with service users; again, this could be valuable evidence to generate, but is not a substitute for evidence of whether the intervention actually delivers benefits.

Comparative studies of similar interventions

Consideration needs to be given to the mechanisms that would ensure comparative studies are undertaken for similar interventions, potentially including addressing which work best for which sub-populations. Given the less-competitive nature of housing providers (compared to drug companies), the ability to collaborate to identify the best approaches that all partners could subsequently deploy seems likely to be relatively high.

Common currencies

In seeking to compare impacts on different outcomes on a consistent basis, the adoption of a common currency can provide a useful tool. In health the preeminent common currency is the quality-adjusted life year (QALY): an intervention that delivers one QALY might be extending healthy life by one year or improving someone's quality of life from 60% to 80% for five years. The adoption of QALYs requires robust evidence of both the impact of an intervention on health outcomes, and the relative reduction in life quality that people experience when living with a condition, and for these to be combined to make an estimate of the QALY impact.

Other sectors have also employed common currencies, for example the measure contained within the Sutton Trust/EEF Education Toolkit that indicates the impact on pupils' attainment in terms of the equivalent number of additional months of schooling that the intervention represents.

Use of tools with evidence of validity

The use, where appropriate, of existing tools for outcome measures that have accrued evidence of validity can avoid risks related to poor design of measures and may facilitate easy translation to other measures that are of interest to potential users of the research. For example, the EQ5D tool for assessing quality of life has accrued evidence of validity across several populations and domains, and would make evidence more useful in terms of conversion to QALYs where a potential evidence user is in the health sector.

Service-user and public involvement

In health research, patient and public involvement in research (PPI) is becoming increasingly common. Involvement in research is distinct from participation, and refers to patients being involved during the design and implementation of the research, not as subjects of the research. PPI is thought to lead to treatments that better meet people's needs and are more likely to be put into practice.[23]

[23] http://www.ucl.ac.uk/jro/patient-public-involvement

Patients may have different aspirations and thoughts about health outcomes that researchers may not have considered.[24] Involving service users early in the design of the research could lower the risk of conducting research that fails to address outcomes that are important for service users.

Service user involvement should also help to avoid the possibility of designing a research project that is difficult for service users to participate in (which might include practical concerns or cultural sensitivities), which would result in recruitment and retention difficulties.

Publication of underlying data to allow reproduction of findings

Where possible and appropriate, the publication of underlying data (suitably anonymised) can improve the overall quality of the evidence base by making it possible for other researchers to reproduce the findings and check for errors.

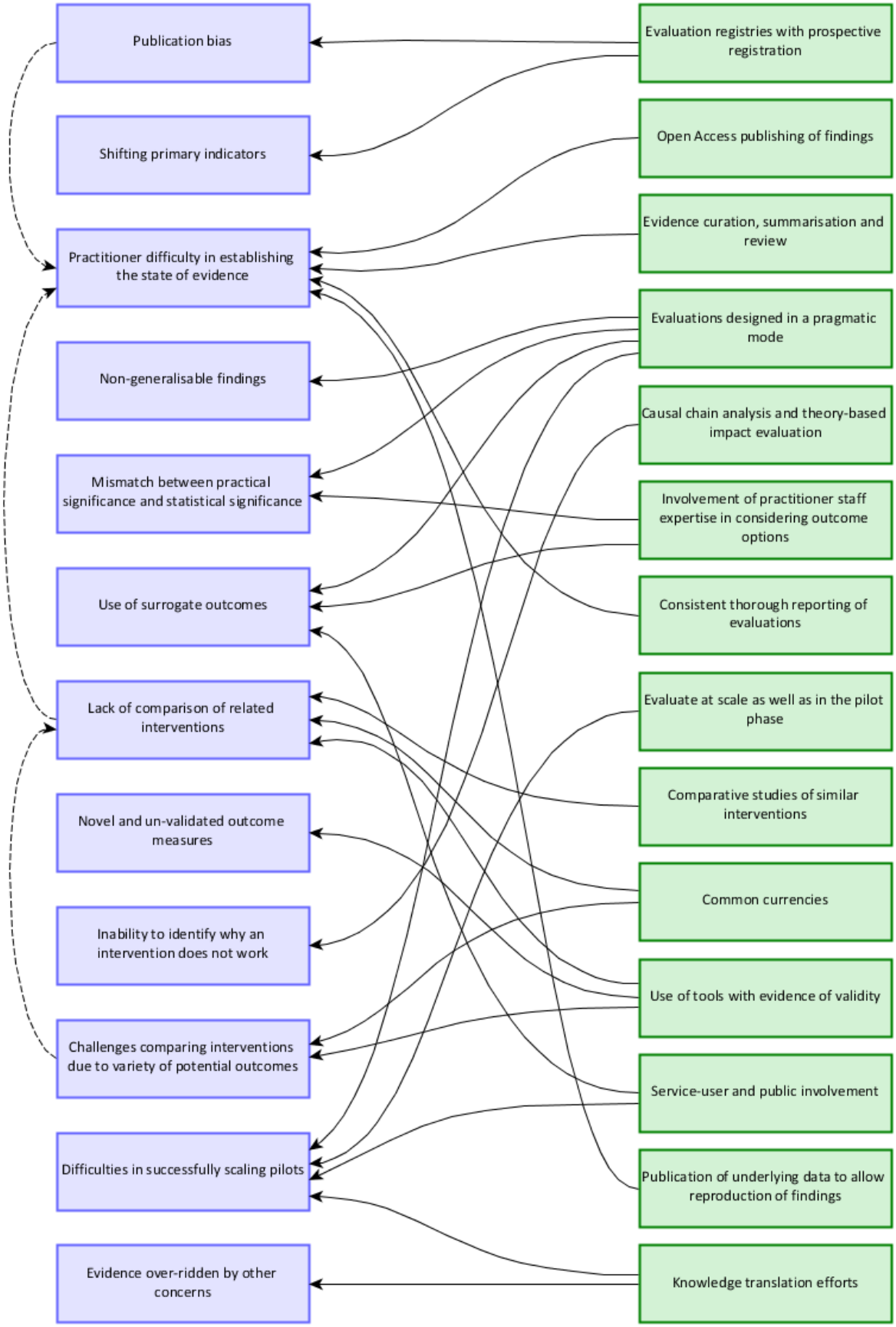
As the state of the evidence base develops to the stage where meta-analyses (studies based on synthesis of existing quantitative research evidence) are possible, the availability of data may also contribute to the ease with which findings from different studies can be combined.

Knowledge translation efforts

In a democracy decision making will never be entirely removed from political concerns. However, effort invested in considering how evidence will be fed into practice will help to create an environment where evidence has the best chance of informing decisions. Further work will be needed to consider the systems and cultural changes that will best support the translation of knowledge of evidence in the housing sector.

[24] NIHR Research Design Service (n.d.) Patient and public involvement in health and social care research: A handbook for researchers. http://www.rds.nihr.ac.uk/wp-content/uploads/RDS-PPI-Handbook-2014-v8-FINAL.pdf

4. Mapping the problems to the solutions



5. Ethics

Subjecting people to research where there is genuine uncertainty about the potential outcome, and the research has a reasonable prospect of resolving that uncertainty, is likely to be ethically sound as long as the intervention does not carry an undue risk of causing harm. However, it is potentially more problematic to research on people in a way that does not contribute as much to the evidence base as it could. For housing providers, as organisations dedicated to delivering a social mission, there is also an additional ethical dimension regarding making effective use of resources that should be delivering public benefits.

Consequently, it could be argued that there is an implicit ethical motivation to the resolution or avoidance of many, if not all, of the problems highlighted in this paper: whether it is allowing research to go unpublished leading to others wasting time duplicating it or conducting an overpowered study that means fewer potential interventions are rigorously assessed than could have been, establishing systems that best enable evidence to be created in a way that will meaningfully contribute to better services from housing providers is a positive societal good.

This does not mean that we should be paralysed for fear of creating evidence incorrectly; nor does it mean that only some mythical 'perfect' study would be ethically appropriate; nor that every study has to establish the definitive facts on a subject with 100% certainty. For example, even an underpowered study can contribute to broader knowledge if it is conducted in such a way that data could be included in later analysis. However, the goal of research should always be to create robust evidence that advances knowledge of a subject. Poor quality research risks wasting scarce resources and misdirecting service delivery.