

In our recent paper ‘Trust and the Goldacre Review: Why TREs are not about trust’[1] we argue that Trusted Research Environments (TREs) reduce the need for trust in the use and sharing of health data, and that referring to these data storage systems as ‘trusted’ raises a number of concerns. Recent replies to our paper have raised several objections to this argument. In this reply, we seek to build on the arguments presented in our original paper, address some of the misunderstanding of our position expressed in these replies, and sketch out where further research is needed.

### *Trustworthiness Signalling*

One of the central arguments of our original paper was that the language we use to describe data-sharing initiatives matters, because the way things like TREs are presented and framed makes a difference to what people can, and should, expect from them. However, there is a more basic and obvious problem with referring to these institutions as ‘trusted’: it begs the question about whether these institutions are, in fact, trusted. We cannot know in advance of these initiatives being implemented, established, and appropriate assessments completed, whether they are actually trusted. Further questions abound: whom might TREs eventually be trusted by? If some people end up trusting them, but not others, is it appropriate to call them ‘trusted’? At best, this is simply a case of wishful thinking (and a more accurate name would be ‘hopefully-trusted research environments’). More problematically, this may be an instance of ‘trustworthiness-signalling’: an attempt to convince others to trust, without meeting the requirements for genuine trustworthiness. By calling TREs ‘trusted’, the implication seems to be that placing our trust in them is perfectly fine; they are the kinds of things that are trusted by others and can safely be trusted by us. But not only does simply calling something ‘trusted’ not make it so, calling something ‘trusted’ does not make it trustworthy. Even if TREs *were* trusted in some meaningful sense, it would not follow that this trust is warranted.

In their commentary to our paper, Affleck et al. [2] describe several initiatives designed to investigate what features would make health data sharing trusted by the public. These include things like security, a lack of commercial involvement, and measures to ensure transparency and accountability. Because these are the sorts of features that TREs provide, Affleck et al. argue, TREs can help to make the system for data sharing more trusted. [2] Again, this remains an assumption until TREs are implemented and we see whether they are so. More importantly, convincing the public that TREs can be safely trusted (i.e., that they are *trustworthy*) seems to be the reason for describing them as ‘trusted’ in the first place. Yet, even if TREs signal that they satisfy all of the conditions that public consultations suggest that they should, this is not sufficient to make them trustworthy. Nor of course, is it sufficient for them to actually be trusted.

There are costs to ‘trustworthiness-signalling’. When we place our trust, we take on the risk of having our trust betrayed. Contrast this with a related concept: reliance. To rely on someone (or something) to X is simply to act on the supposition that X will happen. [3] We often rely on inanimate objects (e.g., alarm clocks to sound, car engines to start), but can also rely on people (e.g., that the coffee shop won’t close early, that my partner will bring home dinner). If my car doesn’t start, or the coffee shop closes early, I might be disappointed or upset, but it would be inappropriate to feel betrayed in the way that I would if my trust had been let down (e.g., if the coffee shop owner was my friend, and she had promised to stay open for me). As we argue in our original paper, if we invite others to trust (which calling something ‘trusted’ is evidently meant to do), we take on the additional normative weight that comes with being trusted rather than merely relied on. People’s expectations will adjust based on what is expected of a trust relationship, as will the consequences of failure. If an institution is inviting the public to trust, they have an obligation to be trustworthy. Where institutions are unable to fulfil this obligation, they risk being guilty of betrayal (rather than simply disappointment), which raises the stakes associated with data sharing.

One of the main objections to our original paper concerned the argument that TREs reduce the need for trust in health data sharing. As multiple commentators have pointed out, TREs are not infallible, nor do they determine all possible uses of health data. Researchers, data controllers, and those associated with the TRE itself will each still need to be trusted to a certain extent. We agree. Our claim is that because TREs are intended to reduce uncertainty associated with health data sharing, they reduce the space in which trust is required as a response to that uncertainty. Moreover, the way that TREs strive to reduce uncertainty is by constraining certain kinds of behaviours, which we take to be an attempt to make health data sharing more reliable, rather than more trustworthy. As we state in our original paper, trust is not merely a prediction about others' behaviour, and thus, trustworthiness is not mere predictability. Accordingly, taking steps to make health data sharing more predictable by constraining the ways that it can be used is not enough to make the institutions collecting, sharing, and using it, trustworthy.

### *'Building Trust'*

In their commentary, Affleck et al. argue that because TREs reduce uncertainty, "they allow trust to grow, rather than diminish" [2]. Similarly, Jesudason argues that verification can reduce uncertainty, but nevertheless increase "feelings of trust" [4]. Part of the disagreement over how TREs do or do not involve trust likely stems from the fact that the notion of 'building trust' is somewhat ambiguous, so it is worthwhile sorting out exactly what it means to 'build trust' in the context of sharing health data, and its relationship to security and monitoring.

There are two ways of understanding the claim that TREs help to 'build trust' in health data sharing: as a descriptive claim, or as a normative claim. As a descriptive claim, the idea seems to be that because the security and monitoring measures offered by TREs reduce the risk of data misuse, the public is (or will be) more willing to trust data sharing initiatives that use TREs. The empirical literature cited by Affleck et al. seems to support this: members of the public report that among other factors, increased security, monitoring, transparency, and accountability make them more willing to trust data-sharing initiatives [2].

However, one might object that people often conflate trust and reliance, and that empirical measures of 'trust' might actually be describing something else. Indeed, our ordinary usage of the word 'trust' often lacks the precision which we might think, on reflection, is important (e.g., 'I'm not sure I trust that ladder', compared to 'I trust my doctor'). The extent to which measures of trust capture 'trust' in this more precise sense is also subject to considerable debate [5], and may be associated with differing sets of attitudes in different contexts [6]. As we argue in our original paper, factors like enhanced security and monitoring seem not to be concerned with trust in the sense in which we might trust those close to us, insofar as they seem to be about making data sharing more predictable. Rather, they seem more concerned with reliance. This does not mean that these factors are any less important a part of an ethical system of health data sharing, but it does highlight the need for conceptual clarity in evaluating public attitudes about data sharing. When we suggested in our original paper that it might be time to move beyond talk of trust, we did not mean that factors like data security are unimportant, but rather that perhaps it would be useful to frame our relationship to those using, sharing, and storing health data as primarily one of reliance, rather than trust.

Moreover, even if empirical measures of trust do capture 'trust' in the more precise sense, we might question the significance of the descriptive claim. Why is it important that the public trust health data sharing in this way? People regularly trust when they should not, and fail to trust when they should, so it is not clear that trust is a reliable indicator of trustworthiness. We might 'build trust' in

health data sharing by misleading or manipulating the public in various ways (at least until this manipulation or misleading is discovered), just as we might build it through improved security, transparency, or accountability. Indeed, Jesudason emphasizes the wide range of social factors that influence feelings of trust [4]. While in some cases the influence of these social factors may be well-founded (e.g., distrust in the medical establishment by historically marginalized groups), in other cases they are not (e.g., trusting a product because a celebrity endorses it). As we argued above, because trust is not always an accurate indicator of trustworthiness, the fact that health data sharing is trusted by the public or by some section of the public does not mean that this trust is warranted. If public trust in health data sharing (or at least certain aspects of it) is important—and it seems likely that it is—its importance is contingent on it being well-placed.

Along these lines, we might understand the idea that TREs contribute to ‘building trust’ in the system as a normative claim about moving towards well-founded trust, or trustworthiness. This is perhaps what Affleck et al. have in mind when they conclude that TREs “provide evidence of trustworthiness in the use of our healthcare data” [2]. For example, in our original paper, we give an example of a babysitter being monitored through a hidden camera, arguing that in this case the parent is failing to trust. Conversely, Affleck et al. claim that further evidence gathering about the babysitter’s activities allows trust to grow rather than diminish [2]. It is possible that what is occurring is the parent is gathering evidence of the babysitter’s trustworthiness, just as the security and monitoring offered by TREs provides evidence that we are building a system that is trustworthy. Presumably, in the babysitter case, the parent would stop monitoring the babysitter once they had gathered enough evidence of the babysitter’s trustworthiness – with such evidence, the parent would then be in a position to trust. However, TREs are not designed to operate in this way, with security and monitoring being reduced over time. This suggests that the security and monitoring offered by TREs is not meant to provide evidence of the trustworthiness of TREs, at least with respect to those facets of data-sharing for which TREs provide security and monitoring.

Alternatively, we might think that the use of TREs is not meant to establish that TREs themselves are trustworthy, but rather that the institutions that make use of them are trustworthy by providing assurances of certain behaviours (i.e., of reducing what needs to be trusted). Does constraining one’s behaviour, even voluntarily, make one more trustworthy? Return to the babysitter example above. Suppose that the babysitter is very attentive to the child, except when she gets distracted watching television. Rather than have the parent trust that she will avoid watching television, so as not to get distracted, the babysitter offers to ‘lock out’ the television. Suppose this makes the babysitter more trusted by the parent. Does it also make her more trustworthy? It does not seem to make her more trustworthy with respect to television-related distractions; rather, it eliminates them as a matter of trust. But what about her trustworthiness ‘as a babysitter’? Does it make the parent’s trust more or less warranted (i.e., does it imply anything about how the babysitter will behave in future instances where trust might be required?) On the one hand, it suggests that the babysitter may simply avoid the need for trust; she may be reliable but not necessarily trustworthy. Or, perhaps it is an indication that the babysitter values the proper care of the child, and is willing to take the necessary steps to ensure this. If this is the case, the babysitter does seem to be demonstrating her trustworthiness. On the other hand, it is possible that taking steps to avoid the possibility of distraction is simply a means of keeping the babysitting job and is not an indication that the babysitter will behave in a trustworthy way in the future.

Accordingly, the use of TREs does not necessarily makes the institutions that use them more trustworthy. It is clear that they reduce the scope of where trust is required with respect to data sharing; it is less clear that this demonstrates their trustworthiness. Perhaps trustworthy institutions

recognize where they should prioritize reliability rather than seeking to be trusted, and consequently where they limit or foster the potential for discretionary action.

A final possibility, alluded to above, is perhaps the most plausible: TREs are not about trust themselves but are part of a system which is, overall, aiming at being trustworthy. We can understand this system in broad terms, as the overall set of institutions which are involved in the access, sharing and use of patient-level health data for research. This will include governance and oversight, ethics and regulatory committees, research institutions and healthcare institutions. We might have good reason to think that trustworthiness matters in this broader context and is one in which data security will be very important. TREs, as secure facilities for data, will be crucial in securing this overall, system level trustworthiness but their role in this system is to be secure and provide guarantees – not to be trusted or trustworthy. Indeed, we have argued elsewhere that in cases where consistent and guaranteed performance is required, assurances of reliability may be preferable to trustworthiness [7,8]. As we suggested in our original paper, much in the Goldacre Review aims at these system level processes and their efficiency which we think is definitely worth pursuing. This fits nicely with the claim by Affleck et al about growing trust – not trust in TREs but in the overall system.

In summary, TREs are (still) not about trust, although they may be part of building a health data research system that is trustworthy. What this discussion illustrates is one of the central arguments of our original paper: we need to be clear about what we mean when we talk about ‘building public trust’ in health data sharing. We need clarity about what institutions mean when they talk about trust and trustworthiness. We need clarity about what the public mean when they talk about trust and trustworthiness. This needs work and further examination.

## References

- 1 Graham M, Milne R, Fitzsimmons P, Sheehan M. Trust and the Goldacre Review: Why trusted research environments are not about trust. *J Med Ethics* 2022. doi: 10.1136/jme-2022-108435.
- 2 Affleck P, Westaway J, Smith M, Schrecker G. Trusted research environments are definitely about trust. *J Med Ethics* 2022. doi: 10.1136/jme-2022-108678.
- 3 Holton R. Deciding to trust, coming to believe. *Australasian Journal of Philosophy* 1994;72(1):63-76.
- 4 Jesudason E. Verification and trust in healthcare. *J Med Ethics* 2022. doi: 10.1136/jme-2022-108634.
- 5 Organisation for Economic Cooperation and Development. OECD Guidelines on Measuring Trust. OECD Publishing: Paris, 2017. Available <https://www.oecd.org/governance/oecd-guidelines-on-measuring-trust-9789264278219-en.htm>.
- 6 Sheikh Z, Hoeyer K. “That is why I have trust”: Unpacking what ‘trust’ means to participants in international genetic research in Pakistan and Denmark. *Med Health Care Philos* 2018;21(2):169-179.
- 7 Sheehan M, Friesen P, Balmer A, et al. Trust, trustworthiness, and sharing patient data for research. *J Med Ethics* 2021;47:e26.
- 8 Graham M. Data for sale: trust, confidence, and sharing health data with commercial companies. *J Med Ethics* 2021; doi: 10.1136/medethics-2021-107464.