Time, space and form: Necessary for causation in health, disease and intervention?

David W. Evans¹ · Nicholas Lucas² · Roger Kerry³

Abstract Sir Austin Bradford Hill’s ‘aspects of causation’ represent some of the most influential thoughts on the subject of proximate causation in health and disease. Hill compiled a list of features that, when present and known, indicate an increasing likelihood that exposure to a factor causes—or contributes to the causation of—a disease. The items of Hill’s list were not labelled ‘criteria’, as this would have inferred every item being necessary for causation. Hence, criteria that are necessary for causation in health, disease and intervention processes, whether known, knowable, or not, remain undetermined and deserve exploration. To move beyond this position, this paper aims to explore factors that are necessary in the constitution of causative relationships between health, disease processes, and intervention. To this end, disease is viewed as a causative pathway through the often overlapping stages of aetiology, pathology and pathophysiology. Intervention is viewed as a second, independent causative pathway, capable of causing changes in health for benefit or harm. For the natural course of a disease pathway to change, we argue that intervention must not only occupy the same time and space, but must also share a common form; the point at which the two pathways converge and interact. This improved conceptualisation may be used to facilitate the interpretation of clinical observations and inform future research, particularly enabling predictions of the mechanistic relationship between health, disease and intervention.

Keywords Bradford Hill · Causation · Disease · Evidence-based medicine · Health · Intervention · Mechanism · Treatment

Introduction

Sir Austin Bradford Hill’s ‘aspects of causation’ (Hill 1965) arguably represent the most influential thoughts on the subject of proximate causation in health and disease, since the postulates of Koch (1884) facilitated identification of microbial causative agents. Hill compiled a list of features (Table 1) that, when present and known, indicate an increasing likelihood that exposure to a factor causes—or contributes to the causation of—a disease.

This list had value because the correlation between two or more variables is always a fallible indicator of causation, and is subject in particular to the problem of confounding (Clarke et al. 2013). On introducing his list, Hill pondered the following problem: “Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?” Howick et al. (2009) recognised that Hill’s list was not only applicable to proximate aetiology in disease causation, but could also be applied to causation in more general terms, including therapeutic intervention, and so modified the list accordingly (Table 1).
Hill deliberately avoided describing the items of his list as ‘criteria’, as this would have inferred that every item was necessary for causation. Instead, his focus was on factors that were known or at least knowable, so that a conclusion could be drawn about the probability of causality. Hence, criteria that are necessary for causation in health, disease and intervention processes, whether known, knowable, or not, remain undetermined and deserve exploration.

It is evident that the items on Hill’s list operate over two ‘natural systems’ (Engel 1980): the individual person, and the population in which that person exists. Studies of health and disease in the latter are important for gaining estimates of effect size in a probabilistic sense. However, such studies are not infallible. Effect size can only be accurately estimated if the population being studied contains sufficient numbers of people with the same disease (i.e. the population is fairly homogenous with a high prevalence of the disease). Hence, the effects of an intervention are only likely to be measurable in a population with a high prevalence of a tractable disease and not in a population without such prevalence.

Such population studies add little to an understanding of how interventions act in a mechanistic sense. It is therefore only conceivable that factors independently operating at the level of an individual could be necessary for a causative relationship between health, disease and intervention; contenders from Hill’s list are thus limited to temporal proximity (time), spatial proximity (space) and the acting mechanism itself (form). The necessity of each is argued in detail below.

### Table 1  Modified list of Bradford Hill for proximate causation in relation to health

<table>
<thead>
<tr>
<th>Item</th>
<th>Necessary</th>
<th>Explanation</th>
<th>System of operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal proximity</td>
<td>Yes</td>
<td>The outcome must occur after or during exposure to the active causative factor, never before</td>
<td>Individual</td>
</tr>
<tr>
<td>Spatial proximity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes</td>
<td>Causative factors must be able to act upon the individual and eventually the site(s) of outcome</td>
<td>Individual</td>
</tr>
<tr>
<td>Mechanism</td>
<td>Yes</td>
<td>A mechanism of interaction between the causative factor and the biology of the recipient must exist. However, it is not necessary to know the details of the mechanism to measure its effect</td>
<td>Individual</td>
</tr>
<tr>
<td>Coherence</td>
<td>No</td>
<td>Cause-and-effect interpretations should not seriously conflict with known facts of the condition and causative factor. Generally, the evidence from basic science and population studies should be mutually supportive</td>
<td>Individual and population</td>
</tr>
<tr>
<td>Specificity</td>
<td>No</td>
<td>To infer a causal relationship, it helps if the effect has only one likely cause. However, in reality outcomes often have multiple causative factors, and exposures can simultaneously cause more than one effect</td>
<td>Individual and population</td>
</tr>
<tr>
<td>Biological gradient (Dose-responsiveness)</td>
<td>No</td>
<td>Greater exposure should generally lead to greater incidence of the effect, if the purported mechanism predicts such a relationship. Strongest ‘dose–response’ evidence comes when the process is reversible; when reduced exposure reverses the effect</td>
<td>Individual and population</td>
</tr>
<tr>
<td>Strength</td>
<td>No</td>
<td>The larger the association between exposure and outcome, or the larger the ratio of an exposed group versus an unexposed group, the more likely that the relationship is causative</td>
<td>Population</td>
</tr>
<tr>
<td>Experiment</td>
<td>No</td>
<td>The size of observed effect is not likely attributable to confounding</td>
<td>Population</td>
</tr>
<tr>
<td>Replicability</td>
<td>No</td>
<td>Replicability of research results, faithfully repeated and consistently observed by different persons in different places, circumstances and times, with different samples, strengthen the likelihood of a causal relationship</td>
<td>Population</td>
</tr>
<tr>
<td>Similarity</td>
<td>No</td>
<td>Similar effects of observed factors in similar populations may be considered as strengthening the likelihood of causation</td>
<td>Population</td>
</tr>
</tbody>
</table>

<sup>a</sup> Did not appear in Hill’s original list (added by Howick et al. 2009)

### The bare necessities

Hill deliberately avoided describing the items of his list as ‘criteria’, as this would have inferred that every item was necessary for causation. Instead, his focus was on factors that were known or at least knowable, so that a conclusion could be drawn about the probability of causality. Hence, criteria that are necessary for causation in health, disease and intervention processes, whether known, knowable, or not, remain undetermined and deserve exploration.

It is evident that the items on Hill’s list operate over two ‘natural systems’ (Engel 1980): the individual person, and the population in which that person exists. Studies of health and disease in the latter are important for gaining estimates of effect size in a probabilistic sense. However, such studies are not unproblematic. There are, for example, known issues in extrapolating probability judgements drawn from population data to individuals (Subramanian et al. 2009). Further, although experimental designs (as utilised in randomized trials) set out to reduce confounding, their results are not infallible (Clarke et al. 2013). Effect size can only be accurately estimated if the population being studied contains sufficient numbers of people with the same disease (i.e. the population is fairly homogenous with a high prevalence of the disease). Hence, the effects of an intervention are only likely to be measurable in a population with a high prevalence of a tractable disease and not in a population without such prevalence.

Such population studies add little to an understanding of how interventions act in a mechanistic sense. It is therefore only conceivable that factors independently operating at the level of an individual could be necessary for a causative relationship between health, disease and intervention; contenders from Hill’s list are thus limited to temporal proximity (time), spatial proximity (space) and the acting mechanism itself (form). The necessity of each is argued in detail below.

### Time

Until now, in accordance with a Humean notion of causation, temporal proximity has been considered the only item from Hill’s list to be necessary for causation. It is evident that this necessity can operate at the level of an
individual person, entirely independent of the wider population. In terms of change in the health of a person, this applies equally to disease aetiology (Russo and Williamson 2007) and therapeutic intervention (Howick et al. 2009); exposure to aetiological agent or intervention must precede, or occur simultaneous to, the change in health. The time interval between the exposure and expression of the outcome should be in accord with the putative mechanism of action. From a probabilistic viewpoint, the shorter the interval, the less room exists for confounders, especially spontaneous remission, to interfere with measurement of effect (Howick et al. 2009).

The timing of intervention application also has mechanistic importance. Therapeutic efficacy depends upon simultaneity of the action of the intervention with the targeted stage of a disease being ‘active’, and therefore susceptible to the influence of intervention. For example, it is pointless applying a preventative measure such as opening a parachute on a person falling from a great height (proximate aetiological factor) once they have reached the ground. Similar scenarios apply to the use of motorcycle helmets or car seat belts in a road traffic collision, gas masks or biohazard suits in the presence of a harmful agent, birth control once pregnant, vaccination once infected by the pathogen that served as the vaccine template, and so forth. As for treating pathology in progress, good examples are the timely repair of an aortic aneurysm, or resection of a primary cancer before metastasis.

**Space**

Howick et al. (2009) astutely added spatial proximity to Hill’s list, although they did not explicitly consider it necessary in the constitution of a causative relationship, despite it acting at the level of an individual person. We would go further and argue that it is necessary when interpreted such that proximate aetiological factors must be able to act upon the individual. This gives a dimension of spatial dependency to any potential causative agent, harmful or beneficial.

One may consider exposure to external agents such as the immediate environment through the respiratory organs, mucosa or skin, infectious or parasitical agents carried by vectors. A force that exceeds the failure limit of body tissues must act on these tissues to sustain injury. Further, such external agents must at some point gain access to the eventual site(s) of outcome. For example, a cut from a dirty knife can introduce a microorganism into an environment where it can thrive, reproduce freely, and in doing so harm its host. Certain strains of *E. coli* are beneficial in one part of the gastrointestinal system but not another; they are even more harmful should they get into the bloodstream or across the blood–brain barrier. Smoke is most dangerous to a person when inhaled into the lungs. Even an external ‘trigger’ for a psychological disorder can gain access through the senses (e.g. leading to post-traumatic stress disorder).

The mere presence or absence of a factor within a person appears to suffice in terms of spatial dependency (since absence can still constitute a crisis within a person); the lack of a factor essential to a person, such as air, water or food will eventually result in a serious adverse effect on health. Not all disease, however, is caused by the presence or absence of external agents.

We are born with our genes, which provide some level of health determinism throughout life. Genes do not exist to cause diseases though (Ridley 1999); aside from reproduction, DNA exists to transcribe RNA, which in turn is coded for the manufacture of protein (Crick 1958). Nevertheless, there are many examples of health conditions that result from the spatial presence of faulty, absent or superfluous genes in our cells. One single extra or missing base in a sequence of DNA can shift the way the sequence is translated—a ‘reading-frameshift’ mutation—resulting in a very different protein structure. An extra or missing complete gene can also threaten life. Moreover, with the exception of chromosome 21, having an extra copy of a whole chromosome is incompatible with life beyond a few days post-partum. Even so, those with the resulting ‘trisomy 21’ will suffer delayed growth, be intellectually less able, and live a relatively short life.

As with disease aetiology, we also view spatial proximity as a mechanistic necessity in the application of therapeutic intervention, primarily because it must also be able to access and act upon the person. It seems nonsensical and irrational that treatment can be applied in the total absence of the patient. Any form of surgery serves as a good example for this assertion. The advent of digital telecommunications has meant that it is now possible to deliver some treatments remotely. Robotic surgery, exercise instruction, education and counselling may all be delivered with the practitioner theoretically being situated anywhere, the only requirement being the presence of reliable telecommunications.

The minimum requirement for an intervention to act from a distance is that the recipient is sensorily accessible, typically visually or aurally. Indeed, sensory stimulation serves as a good mechanistic example of spatially proximate interventions in the arena of pain management. Topical sensory stimulation may be used for therapeutic gain: transcutaneous electrical nerve stimulation, acupuncture, electro-acupuncture, and various ointments (e.g. capsaicin, the active ingredient from chilli peppers). As with all approaches to pain relief, the form of the interaction between intervention and ongoing biology must take place within the nervous system. The special senses
may also be used to gain access to the central nervous system to effect pain reduction, for example through the use of mirrors with amputees suffering from phantom limb pain to create the visual illusion of an intact limb (Chan et al. 2007). In a similar way, phantom limb pain can be temporarily improved by stimulating the vestibular apparatus, the primary organ providing balance information, by the rapid introduction of cold water to the ear canal (André et al. 2001).

The intervention or its resultant processes (e.g. drug metabolites) must be capable of reaching the site(s) where the disease process is active. Hence, the potential efficacy of an intervention depends upon the precision of delivery to such a target. An illustration of this would be the pointless application of a topical antibiotic cream for pyelonephritis (bacterial infection of the kidney). In this regard, accuracy may be defined as the intervention, or a derivative of it, acting upon the intended target, and specificity is the intervention not acting upon other areas (Ross et al. 2004).

As an example, the validity of diagnostic local anaesthetic injection ‘blocks’ depends entirely upon target accuracy and specificity (Engel et al. 2014). Here, accuracy equates to the block succeeding in anaesthetising the target structure, whereas specificity means the block does not anaesthetise other structures that might feasibly be a rival source of pain. In the latter situation, the diagnostic inferences drawn will likely be wrong.

The effect and fate of external substances can also differ according to the route of administration. For example, some drugs can only be absorbed through one route (e.g. oral) whilst others must be delivered through another (e.g. intramuscular). Of prime importance, the incorrect spatial application of an otherwise appropriate treatment at a typically therapeutic dose can be harmful, effectively becoming a proximate aetiological factor for pathology. Consider obvious examples where physical treatments such as surgical techniques or acupuncture are misapplied. Similarly, whilst saline administered intravenously and air pumped into the lungs can both preserve life, each could have the opposite effect if they were to exchange places.

Finally, the antitheses to spatially dependent interventions are remote treatments that also do not interact with the senses. The most obvious example of these approaches is prayer. Several trials have evaluated intercessory prayer as an intervention for life-threatening conditions, using a range of outcomes. Most striking when these results are combined (Roberts et al. 2009) is that prayer has been shown to have no effect on the most important outcome of all, mortality. This is important because one may assume that death is arguably the outcome most undesirable for those who pray to improve the health of others. Far from offering evidence supporting prayer as a therapeutic intervention, one trial (Leibovici 2001) was perfectly designed to demonstrate that positive results can sometimes be nothing more than statistical artefacts by violating the current undisputed requirement of causality, time; the cause must precede, or at least be simultaneous to, the effect.

In this study, volunteers were asked to perform the prayer intervention retroactively, several years after the patients were hospitalised for a bloodstream infection (moreover after many had died in hospital). Mortality was not significantly different in the intervention (prayer) and control groups, again showing that prayer had no ‘effect’ on mortality. However, the length of stay in hospital and duration of fever were significantly shorter in the intervention group. If these findings were due to a genuinely therapeutic effect, then prayer can somehow act not only at a distance and with no measurable mechanism, but also in the past; entirely at odds with the results of every experiment ever conceived by scientists in any field.

**Form**

Hill’s original list focused upon the plausibility of a mechanistic interaction, with other mechanistic aspects, such as biological gradient (dose–response) and coherence with “known facts of the natural history and biology of the disease.” Although these mechanistic aspects operate at a level of an individual person, they have never been considered a necessary feature of a causative relationship. Instead, focus has been given to the extent of knowledge for the underlying mechanism (Howick et al. 2010), rather than what constitutes causation itself. Previous scholars have noted that theoretical models are always limited by the horizon of current knowledge and that acceptance of an association as causal is easier when there is a known theoretical basis for such a conclusion (Howick et al. 2009). However, although self-evident, a mechanism of interaction between the intervention and the biology of the recipient must necessarily exist, whether known or not. Indeed, a causative relationship cannot be sufficiently described without this. Hence, acknowledgement that a mechanism exists, and that this is necessary for causation, is the first step toward observing and understanding the form of interaction between an intervention and a person’s health.

An intervention represents an independent causative pathway that can change the course of a person’s health for better or for worse. A disease process is also a causative pathway that changes a person’s health, generally for worse. Hence, it is only if these two causative pathways exist within the same person, converge in time and space and interact with one another that the otherwise natural course of a disease may be altered. In order to successfully change such a natural course, a successful intervention, or
some derivative of it, must act upon one or more process necessary for the continuation of a disease. The precise form of a potentially successful intervention is therefore constrained by being a function of the form of the currently active stage(s) of the disease. Specifically, at least one point in both disease and intervention processes must not only occupy the same time and space, but also share a common form (Table 2).

The convergence of the two causative pathways—disease and intervention—creates a point of singularity with the final stage of each pathway sharing a common form (Table 2). This common form represents the point at which these pathways interact. Theoretically, the form of each pathway could be interpreted at every sub-ordinate system (Engel 1980) down to a scale of the fundamental physical interactions: gravitational, electromagnetic, strong nuclear, and weak nuclear. At some point, both disease and intervention pathways will have the same form. Of course, complex phenomena, such as a phobia and its resolution through graded sensory exposure, are not easily reduced to such fundamental interactions. Hence, until every mechanistic ‘link’ in the causative ‘chain’ (Howick et al. 2010) is known, the highest common form between intervention and disease pathways should suffice for an acceptable understanding of such mechanistic interactions.

It is conceivable that an intervention can harm a recipient. Hence, the direction of change in the course of ongoing disease is important; an interaction of pathways does not guarantee that intervention will provide benefit to the recipient. To do so, whether directly acting upon pathology (e.g. surgical resection), or indirectly through an intermediate mechanism (e.g. hormone therapy), the intervention must play antagonist to one or more disease protagonist. Furthermore, only if an active stage of a disease process is targeted can the course of disease be changed for the better. To prevent disease, the earliest stages in the causative pathway of a disease—aetiological factors—must be reduced or removed before a state of pathology develops. For example, skin damage from ultraviolet (UV) light can be prevented by a material that reflects or absorbs electromagnetic radiation at this wavelength; the common form is the UV light, which is the active stage of the disease pathway.

Likewise, removal of pathology will effectively terminate the causative pathway of disease. For instance, a disease caused by bacterial infection can be terminated by killing or removing the bacteria, or prevented by avoiding exposure to the infectious agent in the first place; the common form is the bacterial colony. Where amelioration is the aim of intervention, the common form will be a countering of a physiological consequence of the pathology, such as replacing insulin in those who can no longer produce it endogenously because of damaged pancreatic cells (diabetes mellitus, type 1).

Once a disease has progressed to a state of pathology, several processes may simultaneously be active. Hence, several processes may be targeted by intervention with multiple intermediate mechanisms operating along a causative pathway from intervention to outcome; in series, in parallel, or both. Mechanistic ‘links’ operating in series are known as ‘mediators’ (Kazdin 2007). For example, clinical trials evaluating angiotensin-converting enzyme (ACE) inhibitors on reduction of stroke mortality might include evidence that ACE inhibitors reduce blood pressure, that reduced blood pressure reduces the risk of stroke, and that the reduced incidence of stroke reduces mortality (Howick et al. 2009).

Elucidating such mediators may enable increased precision when targeting intervention, and wider therapeutic options. When multiple mechanisms concurrently act in parallel to produce an adverse health state, these may be simultaneously targeted by combining treatments. For

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Target process</th>
<th>Common form</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parachute</td>
<td>Person falling from significant height</td>
<td>Force</td>
<td>Reduced velocity</td>
</tr>
<tr>
<td>Sunscreen application</td>
<td>Sunburn</td>
<td>UV light</td>
<td>Reduced skin damage</td>
</tr>
<tr>
<td>Endovascular stent insertion</td>
<td>Arterial aneurysm</td>
<td>Arterial wall</td>
<td>Prevention of arterial rupture</td>
</tr>
<tr>
<td>General anaesthetic</td>
<td>Consciousness</td>
<td>Central nervous system activity</td>
<td>No pain experience</td>
</tr>
<tr>
<td>Stem cell insertion</td>
<td>Damaged tissue</td>
<td>Cells</td>
<td>Functioning tissue</td>
</tr>
<tr>
<td>Surgical resection</td>
<td>Cancer</td>
<td>Malignant cells</td>
<td>Removed pathology</td>
</tr>
<tr>
<td>Antibiotics</td>
<td>Bacterial infection</td>
<td>Bacteria</td>
<td>Removed infective agent</td>
</tr>
<tr>
<td>Resistance training</td>
<td>Muscle weakness</td>
<td>Muscle</td>
<td>Increased motor strength</td>
</tr>
<tr>
<td>Graded exposure</td>
<td>Phobia</td>
<td>Cognition</td>
<td>Reduced sensitivity</td>
</tr>
<tr>
<td>Analgesia</td>
<td>Pain</td>
<td>Somatosensory nervous system activity</td>
<td>Reduced pain</td>
</tr>
<tr>
<td>Insulin injection</td>
<td>Diabetes mellitus (type 1)</td>
<td>Insulin</td>
<td>Reduced blood glucose</td>
</tr>
</tbody>
</table>
instance, general anaesthesia (GA) is commonly used to prevent pain and distress during a surgical procedure by temporarily removing consciousness; a proximate aetiological factor necessary for the experience of pain. However, a GA does not prevent nociception (the neural processes of encoding noxious stimuli) and the potential development of ‘central sensitization’ that may lead to increased post-operative pain (Woolf and Chong 1993). Hence, administration of ‘pre-emptive’ local analgesia (e.g. peripheral nerve block) in the region of surgical intervention, simultaneous to the GA, is now common practice.

One potential threat to the criterion of form as a necessary component of a causative relationship comes from placebo responses. Unlike the effects of prayer, there is little doubt that placebo interventions cause real, measurable effects, and have been shown to alter the course of a wide range of ailments. Indeed, sometimes responses to placebo interventions are larger than those from ‘active’ treatments (Howick et al. 2013).

There is also little doubt that placebo responses are mediated through the central nervous system of each recipient (Benedetti et al. 2005; Moerman and Jonas 2002). This is the spatial dependency of all successful placebo interventions, irrespective of their form. On the other hand, the scope of their form is limited, such that the ailments susceptible to placebo are those that rely on central nervous system processing. In this sense, placebo interventions are unlikely to prevent injury from external force, the consequences of ingesting a poison, or the effects of severe bleeding.

In contrast, pain relief is one of the most commonly studied placebo responses—so called ‘placebo analgesia’. Since activity of the central nervous system is fundamental to the experience of pain (Melzack 1999), it is perhaps unsurprising that functional magnetic resonance imaging (fMRI) studies have shown that particular loci within the recipient’s brain known to associate with painful experiences (including the thalamus, insula, and anterior cingulate cortex) are less active when placebo analgesia is in operation (Wager et al. 2004).

Furthermore, pharmacological experiments have shown placebo analgesia to invariably operate through endogenous opioid pathways within the central nervous system (Benedetti et al. 2005). Hence, placebo analgesia is potentially any stimulus that evokes this pattern of activity, and such stimuli can take several forms. For example, selective reduction of sensitivity to a noxious stimulus can occur in a single body region following the application of topically applied ‘placebo cream’. Voudouris et al. (1985, 1989, 1990) led healthy volunteers to believe that a moisturising cream had analgesic effects by surreptitiously reducing the intensity of an electrical stimulus when stimulating the hand on which cream had been applied. When a higher intensity stimulus was later applied to both hands, one having been ‘treated’ again with the cream, subjects reported a lower intensity of pain in the treated hand.

This site-specific placebo analgesia has since been shown to be mediated by endogenous opioids (Benedetti et al. 1999). This remarkable somatotopic precision suggests that higher brain centres, such as the primary somatosensory cortex that contains the sensory homunculus (or ‘body map’), are involved in placebo analgesia. Consequently, ‘non-specific’ seems to be a poor descriptor for such responses.

**Summary**

For the natural course of a disease pathway to change, we argue that intervention must not only necessarily occupy the same time and space, but must also share a common form; the point at which the two causative pathways converge and interact. The literature on health, disease and intervention is replete with examples where time, space and form are necessary in causative relationships, and there appears to be no obvious counter-examples. Furthermore, each of these factors is mutually dependent; none can lead to a causative event without the others. This improved conceptualisation may be used to facilitate the interpretation of clinical observations and inform future research, particularly enabling predictions of the mechanistic relationship between health, disease and intervention.

**References**


