A unified theory of bone healing and nonunion

BHN THEORY

This article presents a unified clinical theory that links established facts about the physiology of bone and homeostasis, with those involved in the healing of fractures and the development of nonunion. The key to this theory is the concept that the tissue that forms in and around a fracture should be considered a specific functional entity. This ‘bone-healing unit’ produces a physiological response to its biological and mechanical environment, which leads to the normal healing of bone. This tissue responds to mechanical forces and functions according to Wolff’s law, Perren’s strain theory and Frost’s concept of the “mechanostat”. In response to the local mechanical environment, the bone-healing unit normally changes with time, producing different tissues that can tolerate various levels of strain. The normal result is the formation of bone that bridges the fracture – healing by callus. Nonunion occurs when the bone-healing unit fails either due to mechanical or biological problems or a combination of both. In clinical practice, the majority of nonunions are due to mechanical problems with instability, resulting in too much strain at the fracture site. In most nonunions, there is an intact bone-healing unit. We suggest that this maintains its biological potential to heal, but fails to function due to the mechanical conditions. The theory predicts the healing pattern of multifragmentary fractures and the observed morphological characteristics of different nonunions. It suggests that the majority of nonunions will heal if the correct mechanical environment is produced by surgery, without the need for biological adjuncts such as autologous bone graft.

If the reader believes that medicine can only advance with level 1 evidence and prospective randomised controlled trials, it may be best to stop reading now. This article presents a unified theory that links established facts about the physiology of bone and homeostasis,1-2 with those involved in the healing of fractures and the development of nonunion.3,4 This theory was generated by clinicians, for clinicians, after observing the behaviour of bone following a fracture and its treatment. We believe that this theory enhances our clinical thinking by providing a concept to help understand the ways in which a fracture heals, how we influence it as surgeons and how nonunions can be treated.

Theories are a core part of the advancement of science. Although theories represent potential explanations,5 they are not final answers but derived from experience and incomplete evidence. They are guesses – bold conjectures6 that encourage thought and experimentation in order to stimulate the generation of further evidence. We have called this the bone healing and nonunion theory (BHN) and present it to encourage further debate and understanding amongst clinicians.

BHN theory

The key to this theory is the concept that the tissue that forms in and around a fracture should be considered as a specific functional entity. This ‘bone-healing unit’ produces a physiological response to its biological and mechanical environment which leads to the normal healing of bone. Developing after a bone is fractured, the bone-healing unit is active until healing is complete or has failed (nonunion). This tissue responds to mechanical forces and functions according to Wolff’s law,1 Perren’s strain theory3,4 and Frost’s concept of the “mechanostat”.2 Wolff’s law of 18921 describes the physiological response of normal bone to its mechanical environment during growth and remodelling. Perren’s theory of 19783 deals with the physiological response of broken bone to this environment and Frost described bone homeostasis as a “mechanostat” responding to variations in the mechanical environment. The key to understanding these concepts is the...
response of this bone-forming tissue to strain, which is defined as a change in length of a material at a given mechanical load (Fig. 1a). That load may be applied as compression, distraction or rotation. All living biological material is subject to strain, and some tissues respond to it. The strain tolerance of a tissue is the maximum strain at which a tissue will continue to exhibit normal physiological function. Beyond this level, tissues stop functioning normally, or fail. Perren stated that the strain tolerance of lamellar bone is 2%.4

The bone-healing unit normally changes with time, producing different tissues that can tolerate various levels of strain. The granulation tissue that initially forms at the site of a fracture has a strain tolerance of 100%. Subsequently cartilage, with a strain tolerance of 10%, and bone, with a strain tolerance of 2% to 5%, develops. Essentially, the tissue that forms after fracture progressively stiffens the site until the strain is low enough for bone to form. Clearly different parts of the healing mass are exposed to different strains that also change with time. This dynamic load environment leads to the healing and eventual remodelling of the fracture. This is shown in the theoretical graphs in Figures 2 and 3.

**Homeostasis – normal bone physiology**

Under normal physiological conditions, the strain in bone is at a level where it is in homeostasis, with a slow turnover of bone resulting from balanced osteoblast and osteoclast function. This is point B on Figure 2. Following Wolff’s1 and Frost’s2 principles, a prolonged increase in strain will result in increased bone formation, while prolonged reduction in strain results in bone loss. This represents the physiological response and is represented by the area around point B on the hypothetical curve (expanded on Fig. 3). This has been observed in anthropological records of ancient skeletal remains exhibiting hypertrophy linked to activity7 and the bones of modern athletes or those
undertaking heavy manual labour show similar adaptive changes. Beyond this range, there are more extreme effects. A very low strain environment, such as that experienced by astronauts in orbit, is represented by point A on Figure 2 and results in net bone loss. A common, less extreme example is disuse osteopenia following prolonged bed rest or lack of physiological loading. To the right of point B is the zone where active healing occurs.

Experimental evidence
The evidence describing the response of bone to mechanical influences includes data on the cellular response to strain applied in culture, bone healing *in vivo* and finite element analysis. Perren stated that, “the absence of dynamic deformation results in lack of mechanical induction of callus formation” and that, “very small amounts of strain induce callus formation”. Beyond this he stated that, “strain values up to 2% are tolerated by lamellar bone, up to 10% by woven bone but above that (10% to 30%) resorption prevails.” The *in vitro* data suggest that the physiological response range of osteoblasts is at a strain of 1% to 5%, and the strain that stimulates “soft” callus is between 5% and 10%. Clearly the strains which stimulate early (woven) bone healing and those that cause failure of formed (cortical) bone are different. Perhaps the likely position of point B on the BHN strain curve is around 1% and point C at approximately 10%.

A problem with considering the importance of strain is that as a fracture gap shrinks with healing, the strain must increase until logically at some stage, it should become too high for the formation of bone. Nature solves this paradox by lining the osteoblasts up in a spiral, lengthening the final gap, reducing the strain and allowing the formation of bone. We postulate that the observation of this critical biological mechanism explains both how fractures heal and illustrates the importance of strain in this process.

Healing – the normal response to fracture
As part of the normal healing response, haematoma at the fracture site becomes organised and the bone-healing unit forms. Initially, the strain is high (to the right of point C on Figure 2), granulation tissue forms and the healing process gradually stiffens the area until the strain is to the left of point C and bone can form and finally remodel through normal homeostasis. This type of healing is seen after non-

---

**Table I. Biomechanical definitions**

<table>
<thead>
<tr>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relative stability:</strong> A fixation or support construction of a fracture that allows small amounts of motion in proportion to the load applied. This leads to healing by callus formation.</td>
</tr>
<tr>
<td><strong>Absolute stability:</strong> Fixation of fracture fragments so that there is virtually no displacement of the fracture surfaces under physiological load. This allows direct (primary) bone healing.</td>
</tr>
<tr>
<td><strong>Strain:</strong> Change in length of a material when a given force is applied. Normal strain is the ratio of deformation (lengthening or shortening) to original length. It has no dimensions but is often expressed as a percentage.</td>
</tr>
<tr>
<td><strong>Stiffness:</strong> The ability of a material to resist deformation. It is measured as the relationship between load applied and the resulting elastic deformation. The inherent stiffness of a material is expressed by its modulus of elasticity.</td>
</tr>
</tbody>
</table>

---

**Fig. 3**
Graph showing bone homeostasis, healing and nonunion theory (point A to B, Fig. 2) in more detail.
operative treatment of fractures and operative fixation with relative stability (Table I). Clearly the strain is not the same throughout the bone-healing unit and changes during the process of healing.\(^1\) Bone formation probably starts in low strain zones at the periphery before spreading inwards, as strain is progressively reduced in different areas and ossification proceeds across the whole fracture.

**Healing by callus – relative stability**

We hypothesise that there is an optimum strain environment that will maximise the formation of bone with large volumes of callus. From the surgical point of view, this represents healing with “relative stability”. It probably occurs at the higher levels of acceptable strain between points B and C of the response curve shown in Figure 2. It is typically associated with intramedullary nailing.\(^2\)

**Primary bone healing – absolute stability**

Primary bone healing was first described by Danis\(^2\) and occurs when a fracture is reduced anatomically and fixed with absolute stability. This technique was advocated to provide anatomical reduction and sufficient stability in order to allow early mobilisation of the joints.\(^2\) Primary bone healing was a secondary effect. Absolute stability represents a low-strain environment similar to the normal (physiological) strain environment of unfractured bone (i.e., < 2% strain). Healing occurs as a result of normal homeostatic remodelling of the local bone. Healing is slow and there is no callus formation. It corresponds to point B on Figure 2.

**Nonunion – failure of the bone healing unit**

Several authors have described patterns of nonunion by their morphology and callus formation, and suggested that the vascularity of bone was the discriminating factor.\(^3\) Today, the hypertrophic pattern of nonunion is widely accepted to be due to mechanical factors and the atrophic pattern to be avascular. However, experimental models have failed to confirm devascularisation in atrophic nonunion.\(^4\) Indeed, they have shown consistent recovery of vascularity and samples from human atrophic nonunions have been shown to be biologically active, with significant healing cellular activity.\(^5\) While some devascularisation must occur in many injuries, it appears that most fracture fragments are incorporated into healing tissue, surrounded or bypassed by new bone formation, and then become revascularised. Only if infected, critically sized or very displaced, are they likely to contribute to local failure to heal and even in these situations, avascular fragments commonly heal at one end and not the other, suggesting a combined problem of avascularity and local strain concentration. Various other biological factors have also been identified or proposed to affect union,\(^6\) however, their individual role is poorly defined in clinical practice. Today, the damage caused by poorly performed surgery or the failure of poorly applied or incorrect implants are also major factors in the development of nonunions.

We suggest that it may be better to consider nonunions as primarily mechanical or biological in origin, representing the two pathways that can lead to failure of bone healing. Of these, mechanical instability dominates in clinical practice.

In most nonunions, there is an intact bone-healing unit. We suggest that this maintains its biological potential to heal but fails due to the mechanical conditions. This may occur in two biomechanical situations: when the fixation of the fracture or the healing process fails to reduce the strain to a level where healing can occur (remaining to the right of point C on Fig. 2), or if the fixation results in a construct that is so stiff that the strain is always to the left of point B on Figure 2. Examples include persistent movement at the site of a fracture that has not been stabilised (Fig. 4) or following plate fixation of a simple fracture where anatomical reduction is not achieved, leaving a gap with persistently high strain (Fig. 1b). The fracture will fail to heal, and the plate will fail due to fatigue (Fig. 5). The different morphology of nonunions probably represents the spectrum of callus response specific to the local biological response to the mechanical environment.

In situations where very high strain persists, the movement at the fracture site breaks down the bone-healing unit. A ‘synovial’ cavity develops and the bone ends heal, which creates a pseudoarthrosis. In clinical practice, this is seen where an adjacent stiff joint focuses any movement at the site of the fracture, or where there is no inhibition of pain caused by movement of the fracture, as in a neuropathic limb (Figs 4 and 6). In this situation, freshening of the fracture ends and compression osteosynthesis is required to restore the mechanics, recreate the bone healing unit and facilitate union.
Biological influences on bone healing

Many factors have been observed to delay the healing of fractures. These include the severity of the injury, peripheral vascular disease, diabetes, previous irradiation, major endocrine anomalies, infection, smoking and the persistent use of drugs such as non-steroidal anti-inflammatory drugs (NSAIDs). Some of these factors may influence the mechanical response of the bone-healing unit, although the mechanism is unknown. Do they simply delay the formation of bone (Fig. 7, line X), narrow the range of strain at which bone can form (Fig. 7, line Y), or completely prevent bone formation (Fig. 7, line Z)? The biological response may be increased as well as reduced. Enhanced bone formation is seen in some patients with a severe head injury or spinal cord injury and in some cases where bone morphogenetic proteins (BMPs) have been administered. Again, it is not known if this enhanced bone formation takes place over the normal range of strain where bone formation is seen, or is the range of strain, at which bone formation takes place, increased. If the latter, BMPs applied locally to high strain nonunion sites without mechanical stability would be an effective form of treatment. There is, however, no clinical evidence for this at present.

Perhaps the most important additional factor damaging the healing process is poor surgery. This includes poorly controlled dissection and excessive stripping of periosteum causing increased injury to local soft tissues, and the application of stiff plates with a large area of contact. The aim of surgery is atraumatic reduction and fixation of the fracture to allow healing in a better functional position. Poorly executed surgery may delay healing due to devascularisation, or by producing an inappropriate strain environment. The development of minimal footprint plates seems to have facilitated better healing but the major emphasis in fracture surgery must be on meticulous technique that limits additional devascularisation when fractures are exposed and reduced, and the correct application of implants to produce a mechanical environment that facilitates healing.

Stress fractures

Stress fractures occur when repetitive stress applied to a bone results in a focal area of high strain that is above the threshold for normal remodelling of bone, but below the threshold for acute fracture. The failure of bone under repeated stress usually starts on the tension side of the bone. After time, attempts at healing become visible with resorption of bone at the fracture site and bone formation at the periphery. Fracture healing will succeed if the repetitive stress is reduced below the healing threshold, or if the peripheral new bone stiffens the construct to reduce the strain within the fracture site into the healing threshold. If the repetitive stress continues, attempts at healing produce bone but persistence of a strain to the right of point C on Figure 2 leads to failure of bone bridging at the fracture site, and eventually an appearance similar to a hypertrophic nonunion.

A special situation is the atypical fracture associated with the use of bisphosphonates. These drugs uncouple the normal homeostasis of bone. The resulting bone behaves differently and stress fractures occur in some patients exposed to normal physiological levels of strain. In this situation, point C in the strain curve has moved to the left so that failure occurs at a ‘normal’ strain value (Fig. 7, curve Y).

Nonunion in multi-fragmentary fractures

According to Perren, the initial strain environment is relatively low in multi-fragmentary fractures. As the fracture starts to unite, the various fragments heal to each other, eventually leaving a single fracture line with higher strain. If the focal strain remains low enough, uneventful healing will occur. However, if the strain at the last fracture line is above
the threshold for bone formation (to the right of point C, Fig. 2), healing will stop and a single plane nonunion results (Fig. 8). Clinically, nonunion in multi-fragmentary fractures usually presents as a single fracture line with an oblique morphology and multi-fragmentary nonunions seem rare.34 We hypothesise that the obliquity is due to the concentration of the highest local strain produced by shear forces in this plane. If surgical treatment addresses the shear, union should reliably occur.35,36 Accordingly BHN predicts that multi-fragmentary fractures (AO/OTA B or C type diaphyseal fractures)22 that fail to heal will typically produce an oblique single plane nonunion. In general, transverse nonunions are only seen when the original fracture line was either a simple transverse or a stress fracture. Transverse fractures with a small bending-wedge fragment (‘B’ type) should form nonunions following the obliquity of the wedge.

Management of nonunion
If the mechanical environment is critical in fracture healing and dominates the pathogenesis of nonunion, the logical conclusion is that the treatment of a nonunion will be most successful if the mechanical environment is changed to one that facilitates the formation of bone, changing the strain at the nonunion to one that is within the bone formation range (between points B and C, Fig. 2), or one that allows the normal homeostasis of bone (point B). Reducing the strain at a nonunion can be achieved by reducing the forces applied, direct stabilisation or a combination of both. Most simply, it can be reduced by behavioural changes in patients, such as reduced load bearing or limitation of movement by splints. More commonly, stabilisation of the nonunion with a plate, nail or external fixator is performed, reducing the strain enough to facilitate healing. Local strain can be reliably reduced by a device that spans the nonunion and is then preloaded itself. This can be achieved with screws, plates, or external fixators. Shear strain, as is characteristic of an oblique nonunion, clearly must be minimised. This is best achieved by an implant that passes directly through that plane. A low-strain environment can also be achieved by correction of malalignment and with carefully planned fixation with compression plates and lag screws to produce absolute stability. Reliable healing then takes place.35,36 There is no requirement to excise tissue from the fracture site. It will heal if the mechanical environment is corrected.

Autologous bone graft or BMPs are commonly used as biological adjuncts to fracture healing. Despite their frequent use, there is evidence that if a low-strain environment
can be achieved, a nonunion will heal without their use.34,35 Many surgeons perform a Judet decortication as part of the treatment of a nonunion.36,37 This probably creates a supplementary bone-healing unit, indeed elevating the periosseum when introducing bone graft may simply achieve a similar effect. The authors of this paper primarily use mechanical techniques in the surgery of a nonunion and reserve the use of biological adjuncts, including bone graft, to cases where there is significant bone loss.

Many diaphyseal nonunions may be successfully treated with intramedullary nails.38 Most surgeons recommend the use of a nail with a larger diameter than the one in situ. We suggest that increasing the diameter of the nail reduces strain at the site of the nonunion to somewhere between points B and C, allowing bone formation to take place. Clearly, in this situation, it is impossible to separate the mechanical effect of the larger nail from any biological effects of the reaming which is necessary to insert the larger nail. Vascular effects are probably short-lived and after reamed exchange-nailing, successful healing is characterised by callus outside the bone. ‘Autograft’ created by reaming a nonunion cannot be deposited through the nonunion and so cannot contribute to external callus formation suggesting that it is not essential to the healing process.

BHN theory and the Ilizarov technique
Both internal and external fixation techniques can be used to produce stability and facilitate healing. The fundamental requirements of the biology and biomechanics of both are the same – it is only the different ways that the surgeon manipulates these which changes. The Ilizarov method39 of managing a nonunion is consistent with the BHN theory. One of Ilizarov’s key principles was to restore the normal alignment of the limb and then compress the surfaces of the fracture. The 3D construct of a circular frame produces stability, resulting in low strain at the fracture site. The fracture unites as bone formation takes place within the existing tissue.

Increased stability and a reduction in strain can also be produced with distraction of a nonunion or osteotomy. Presumably, distraction at a rate of 1 mm per day creates a strain environment that allows the formation of bone (the area under the line between B and C, Fig. 2) and results in distraction osteogenesis in the bone-forming unit. Greater rates of distraction exceed the strain tolerance and regenerate bone does not form (area to the right of point C, Fig. 2).

Testing the theory
Evidence is needed to support or refute a theory. The BHN theory predicts that there should be specific characteristics of the morphology of a nonunion, cellular biology and the response to treatment that relate to the strain environment, which could be tested by experiments and observation. Specifically, these may include the predicted development of uniplanar nonunions from multi-fragmented fractures, the identification of the biological nature and behaviour of types of tissue by their likely position in our strain curve and the healing of nonunions treated primarily by reduction of strain without additional stimulation.

We have presented a theory that links established views about bone homeostasis, fracture healing and the development of nonunion. It predicts the characteristics of nonunions and considers their surgical management based on a mechanistic view of fracture healing and the assumption that the response is strongly influenced by the mechanical, as well as the biological, environment. Although supported by some experimental and clinical evidence, this is still conjecture. The theory is described in an endeavour to make sense of our clinical experience and to enhance our understanding of this complex process. We hope that it will stimulate debate with challenge and counter-challenge – the essence of science since The Enlightenment.

Author contributions:
R. M. Smith: Wrote paper, Basic concept, Editing, Analysis.
C. G. Moran: Wrote paper, Basic concept, Editing, Analysis.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

This article was primarily edited by J. Scott and first proof edited by G. Scott.

References


