A SURPRISINGLY POOR CORRELATION BETWEEN *IN VITRO* AND *IN VIVO* TESTING OF BIOMATERIALS FOR BONE REGENERATION: RESULTS OF A MULTICENTRE ANALYSIS

G. Hulsart-Billström[1, 2, Ψ], J.I. Dawson[1, Ψ], S. Hofmann[3, 4, 5], R. Müller[3], M.J. Stoddart[6], M. Alini[6], H. Redl[7], A. El Haj[8], R. Brown[9], V. Salih[10, 11], J. Hilborn[12], S. Larsson[2] and R.O.C Oreffo[1]*

[1] Bone and Joint Research Group, Centre for Human Development Stem Cells and Regeneration, Institute of Developmental Sciences, University of Southampton, UK
[2] Department of Surgical Sciences, Orthopedics, Institute of Medicine, Uppsala University, Sweden
[3] Institute for Biomechanics, Swiss Federal Institute of Technology Zurich (ETHZ), Vladimir-Prelog-Weg 3, Zurich, 8093, Switzerland
[4] Department of Biomedical Engineering, Eindhoven University of Technology, PO Box 513, Eindhoven, 5600MB, The Netherlands
[5] Institute for Complex Molecular Systems, Eindhoven University of Technology, PO Box 513, Eindhoven, 5600MB, The Netherlands
[6] AO Research Institute Davos, Davos Platz, Switzerland
[7] Ludwig Boltzmann Institute for Experimental and Clinical Traumatology, AUVA Research Centre, Austrian Cluster for Tissue Regeneration, European Institute of Excellence on Tissue Engineering and Regenerative Medicine Research (Expertissues EEIG) Vienna-Branch, Vienna, Austria
[8] Institute for Science and Technology in Medicine, Keele University, Stoke-on-Trent, UK
[9] University College London, Tissue Repair and Engineering Centre, Institute of Orthopaedics, Stanmore Campus, London, UK
[10] Plymouth University, Peninsula School of Medicine and Dentistry, Plymouth, UK
[11] University College London, Eastman Dental Institute, London, UK
[12] Department of Chemistry, A Science for Life Laboratory, Polymer Chemistry Division, Uppsala University, Uppsala, Sweden

Ψ G. Hulsart-Billström and J.I. Dawson contributed equally to this work and should be regarded as joint first authors

Abstract

New regenerative materials and approaches need to be assessed through reliable and comparable methods for rapid translation to the clinic. There is a considerable need for proven *in vitro* assays that are able to reduce the burden on animal testing, by allowing assessment of biomaterial utility predictive of the results currently obtained through *in vivo* studies. The purpose of this multicentre review was to investigate the correlation between existing *in vitro* results with *in vivo* outcomes observed for a range of biomaterials. Members from the European consortium BioDesign, comprising 8 universities in a European multicentre study, provided data from 36 *in vivo* studies and 47 *in vitro* assays testing 93 different biomaterials. The outcomes of the *in vitro* and *in vivo* experiments were scored according to commonly recognised measures of success relevant to each experiment. The correlation of *in vitro* with *in vivo* scores for each assay alone and in combination was assessed. A surprisingly poor correlation between *in vitro* and *in vivo* assessments of biomaterials was revealed indicating a clear need for further development of relevant *in vitro* assays. There was no significant overall correlation between *in vitro* and *in vivo* outcome. The mean *in vitro* scores revealed a trend of covariance to *in vivo* score with 58 %. The inadequacies of the current *in vitro* assessments highlighted here further stress the need for the development of novel approaches to *in vitro* biomaterial testing and validated pre-clinical pipelines.

Keywords: *in vivo, in vitro*, correlation, biomaterials, multicentre study.

* Address for correspondence:
Professor Richard O.C. Oreffo,
Bone and Joint Research Group,
Centre for Human Development,
Stem Cells and Regeneration,
Human Development and Health,
Institute of Developmental Sciences,
University of Southampton,
Southampton, SO16 6YD, UK
Telephone number: 44 (0)-2381 208502
Fax number: 44-(0)-2381 2085525
Email: roco@soton.ac.uk

Introduction

Researchers and key decision-makers have enthusiastically endorsed the therapeutic promise of regenerative medicine. The anticipated clinical impact has led to considerable research investment and a resultant proliferation of regenerative medicine-related innovations and technologies. One of the most mature research areas in this field is the development of innovative biomaterials for bone regeneration.

Over the last decade, a plethora of increasingly advanced biomaterials, designed to provide specific mechanical or morphological properties and kinetics combined with an expansive range of biomimetic modifications have been developed for use (in isolation or in combination with cell therapy) in bone regeneration (Billstrom *et al.,* 2013; Garcia-Gareta *et al.,* 2015; Gibbs *et al.,* 2014). However, the range of materials and resultant potential therapies present a significant burden for pre-clinical *in vivo* testing and subsequent clinical translation. The challenge for regenerative medicine is analogous to that faced in toxicity testing where the escalating number of new chemical entities (NCEs) and the extensive requirement for animal testing had the result that only a fraction of NCEs were considered for further evaluation. As with NCE toxicity testing, moving away from the use of, or at least reducing, animal testing is a critical imperative in regenerative medicine.

Both ethical and efficiency concerns underlie this imperative. Animal behaviour and neurophysiology appears increasingly to validate the sentiments (not the methods) of the animal rights movement. For example, a study by Bartal *et al.* showed pro-social behaviour in rats (Bartal *et al.,* 2011) and a special vocalisation of adolescent rats is proposed to have an evolutionary relation to the joyfulness of children's laughter (Bering 2012; Panksepp and Burgdorf, 2003). In response to this increasing awareness Russell and Burch's 3Rs initiative is being implemented legislatively and adopted by funders and governments to enable: (1) reduction of the number of animals used, (2) refinement of experiments to minimise animal suffering and distress and (3) replacement of animal testing by alternative *in vitro* approaches (Russell and Burch, 1959).

The extensive fiscal and time expenditure of animal experimentation has provided a further significant driver behind the development for an increased reliance on relatively high throughput *in vitro* alternatives to animal testing. In a systematic review, Morgan *et al.* compared cost estimates of drug developments and observed the cost estimates of preclinical work ranged from USD$ 46 to USD$ 165 billion (Morgan *et al.,* 2011). To date, there remains no gold standard for the estimation of the costs spanning from developing a drug through to clinical implementation.

Successful replacement and reduction of animal procedures has been achieved in various fields, as indicated above and perhaps most notably in toxicology testing, through improvements to *in vitro* and *in silico* models (Langley *et al.,* 2007; Vodovotz *et al.,* 2006) as well as improvements in sharing of existing data in order to inform new studies (Knight, 2008). The same development is necessary in the field of regenerative materials, which needs to be evaluated with reliable and comparable methods. One possibility is to draw conclusions from previously performed experiments/studies through a direct correlation of *in vitro versus in vivo* outcomes of biomaterials. Are there parameters that have been measured within *in vitro* assays that correlate with objective *in vivo* outcomes?

It is important to note *in vitro* and *in vivo* evaluations will likely both remain critically important in the development of a clinical therapeutic entity. *In vivo* experiments allow rigorous assessment of the material-cell/host interaction and the regenerative efficacy of a biomaterial strategy. The wide range of parameters provided through *in vivo* assessment are important for osteo-regenerative materials, as bone itself is a complex organ responsible for protection, muscle attachment, calcium homeostasis and as a centre of haematopoiesis. Bone is a highly organised structure that constantly remodelled and optimised to allow mechanical loading and to meet the demand for calcium by a strictly controlled interplay between osteoblasts and osteoclasts, conducted by osteocytes (Bonewald 2006; Bonewald 2011; Frost 1994). If a defect is not too large, bone has an inherent ability to regenerate completely without scar-formation. This self-regeneration is orchestrated by a number of cytokines and involves a variety of cells where the immune system plays an important part (Marsell and Einhorn 2009; Marsell and Einhorn 2011). Bone regeneration needs certain conditions, where vascularisation, stabilisation, scaffolding, cell signalling and progenitor cells are imperative requirements (Giannoudis *et al.,* 2007; Giannoudis *et al.,* 2008). The implanted biomaterial *in vivo* will encounter the cells of the immune system that will degrade the material or contain the material within a fibrous tissue. Furthermore, the surrounding tissue will provide a mechanical environment that will both affect the material as well as the cell response. The material will have the possibility to affect stem and progenitor cells in terms of lineage fate and function. It is this wide range of parameters that *in vivo* models are designed to encompass. *In vitro* assays, on the other hand, typically, serve as screening assays that allow experimental assessment used as predictive methods for rapid screening of the host response to the biomaterial. The International Organisation for Standardisation (ISO) and ASTM International have developed standards that are specific for *in vitro* biocompatibility and cytotoxicity tests (Müller 2007) and for standardised assessment of biomaterials both *in vitro* and *in vivo* (F1983-14; F2721-09r14), which significantly increases the safety of the product. However, these standards are generally for medical devices and may not predict the bone regeneration of a specific biomaterial. A raft of standardised tissue culture techniques are used to observe cell compatibility where the toxicity of the material can be measured by, for example, cell survival, metabolic activity and cell growth. The effect of the material can be studied utilising cell differentiation where expression of genes and mRNA and cell surface markers are quantified (Ko *et al.,* 2008).

Though *in vivo* experiments will likely remain a critical aspect of biomaterial testing, *in vitro* assays need to play an increasingly important role in screening out biomaterials with inadequate properties prior to examination of more promising candidates *in vivo*. A key question to address is therefore: are current *in vitro* assessments proficient in fulfilling this role of pre-screening prior to *in vivo* assessment of bone regeneration? In this context, the term *in vitro* assessment refers both to the assay itself and the interpreted significance of the assay results that serve to inform further studies along the biomaterial testing pipeline. To address this question, the current study explores the correlation between researcher assessed, *in vitro* and *in vivo* outcomes of biomaterials tailored for bone in a European multicentre study between 8 universities that included 36 independent *in vivo* studies and 47 individual *in vitro* assays testing 93 biomaterial variables. The focus of this study is on the correlation between *in vitro* testing and early stage (*i.e.* small animal) *in vivo* models of bone regeneration in use across the consortium. In contrast to conventional literature based meta-analyses,

Table 1. Participating laboratories in the survey.

University	Group principal investigator	Country
AO Foundation	Martin Stoddart	Switzerland
Eidgenössische Technische Hochschule	Ralph Müller	Switzerland
Keele University	Alicia El Haj	United Kingdom
Ludwig Boltzmann Institute	Heinz Redl	Austria
University College London	Robert Brown	United Kingdom
University of Nottingham	Kevin Shakesheff	United Kingdom
University of Southampton	Richard O. C. Oreffo	United Kingdom
Uppsala University	Sune Larsson, Jöns Hilborn	Sweden

the multicentre approach adopted, though more limited in scope, allowed access to complete data sets. Critically, such datasets included the often unpublished negative data so vital for exploring *in vitro-in vivo* correlations.

Methods

Data collection

Members from the European consortium Biodesign, comprising 8 universities in a European multicentre study, kindly collated the data (Table 1) from both published and unpublished historical *in vivo* datasets. The scope of the data included was defined by completed *in vivo* datasets for the various biomaterial strategies tested. The *in vitro* assays, which constitute the subject of this study, were selected on the basis of their use by the member groups in the testing of their biomaterial strategies prior to or alongside these *in vivo* studies. The data included 36 *in vivo* experiments, 47 *in vitro* experiments and 93 tested biomaterial variables.

In vivo and *in vitro* outcomes for each experimental variable tested in an *in vivo* study were described qualitatively by the assessor on the basis of a range of expected positive and/or negative indications of regenerative outcome (*e.g.* high cell proliferation, evidence of cytotoxicity *etc.)* pre-defined for each assay by the assessor. On the basis of this qualitative assessment, the outcome of each variable for each *in vitro* and *in vivo* assay conducted was scored out of 5 (1 = poor, 5 = very good).

In vitro parameters assessed for correlation

Cell differentiation

Cell differentiation towards the osteogenic lineage can be quantified by measuring the activity of the cell marker alkaline phosphatase (ALP) (McComb *et al.,* 1979), a widely used non-specific assay for osteogenic differentiation. In this assay p-nitrophenyl phosphate is dephosphorylated by ALP into a yellow product with an absorbance that can be measured at a wavelength of 405 nm. ALP is highly expressed by osteoblasts and through the cleavage of pyrophosphate allows spontaneous mineralisation in the secreted osteoid (Millan 2013). A high activity of ALP is interpreted as a positive outcome as ALP plays a central role in the bio-mineralisation process of bone.

Biocompatibility

Several biocompatibility tests can be used for cell proliferation, cell viability, and cell attachment to distinguish material candidates that are of a biocompatible nature. A number of reagents are used and most of them are based on the same concept of measuring the metabolism by the reduction of a substrate. One common example is a tetrazolium compound that is mixed with phenazine methosulphate (MTS) (Goodwin *et al.,* 1995). During cell metabolism, MTS is reduced by the activity of dehydrogenase to formazan that can be measured by visible light absorbance at 492 nm. Live dead staining is a widely used *in vitro* assay to measure the viability of cells. A fluorescent dye 5-chloromethylfluorescein diacetate (CMFDA) labels metabolically active cells and the membrane impermeable ethidium homodimer-1 labels DNA in apoptotic or damaged cells.

Gene expression

Osteogenic differentiation can be measured by quantifying the expression of specific genes expressed during osteoblastic differentiation. One of the key transcription factors is Runt-related transcription factor 2 (RUNX2) (van Wijnen *et al*., 2004), which at early stages induce the process of osteoblast differentiation and in later stages inhibits the same process. RUNX2 is important in skeletal morphogenesis and is involved in the expression of several bone matrix genes such osteocalcin, being an abundant protein found in bone provides useful markers to follow stem-progenitor differentiation. Genes that are expressed at later time points include, for example, osteopontin, involved in remodelling (Bruderer *et al*., 2014). Increased gene expression at the appropriate temporal stages is interpreted as a positive result.

Mineralisation assay

Calcium deposition of cell cultures can be quantified using alizarin red staining and serves as a measure of osteogenic differentiation, where a large amount of deposition is interpreted as a positive outcome (Dawson 1926). This *in vitro* assay provides evaluation of the osteoblastic differentiation of progenitor cells as well as the functionality of the differentiated cell population (Gregory *et al.,* 2004).

In vivo models against which *in vitro* models were correlated

The purpose of this review article was to correlate the interpretation of *in vitro* results to *in vivo* outcomes observed for a range of biomaterials. A selection of *in vivo* models from the survey that had been used to evaluate biomaterials was correlated to the interpreted results of *in vitro* assays for the same biomaterials.

Segmental-defect models

Critical size segmental defects in rats and mice are frequently used to measure the osteoinductive capability of biomaterials for bone regeneration using a fixator that is either internal or external. The definition of a critical sized defect is an incapability of spontaneous healing (Hollinger and Kleinschmidt 1990). Addition of an inductive substance, that triggers bone formation, is necessary for bridging healing to occur. Einhorn *et al*. (1984) developed, in the 1980s, a rat segmental defect that consistently resulted in non-union when left untreated. The femoral defect was 6 mm long and stabilised with two proximal and two distal Steinmann pins. The defect was evaluated with radiographs and mechanical testing (Einhorn *et al.,* 1984). This femoral segmental defect was later reduced to a 5 mm long segmental critical size defect to evaluate the osteogenic potential of silk scaffolds combined with either human bone marrow stromal cells (HBMSCs) or a combination of HBMSCs and bone morphogenetic protein 2 (BMP-2) (Meinel *et al.,* 2006). The same combinations were used with the addition of pre-differentiated HBMSCs (Kirker-Head *et al.,* 2007). Both studies assessed the healing capacity using micro-computed tomography (μCT), mechanical testing and histology.

Both Kaipel *et al.* (2012) and Schutzenberger *et al.* (2012) used a 3 mm long segmental defect as a delayed union model in rats. The defect was stabilised with an internal plate fixator, while initial healing was prevented using a silicone spacer that was kept in the defect for 4 weeks after which it was removed during a second surgery. Following removal of the spacer the defect was filled with either a BMP-2 fibrin carrier or the commercial available BMP-2 collagen carrier. The same defect was used in a similar study to evaluate the effect of applied growth factors of angiogenesis and osteogenesis in a fibrin clot. The bone forming capacity was evaluated by X-ray, μCT and mechanical testing (Kaipel *et al.,* 2012; Schutzenberger *et al.,* 2012). The results showed increased bone formation using BMP-2 in a fibrin clot. The study also showed that a fibrin scaffold with a sevenfold lower dose of BMP-2 could still provide equivalent results compared to the commercial BMP-2 product.

Kanczler *et al.,* used the male MF-1 nu/nu immunodeficient mouse femur segmental defect model to study the enhancement of bone regeneration after implantation of HBMSCs seeded onto vascular endothelial growth factor (VEGF)/BMP-2 composite scaffolds (Kanczler *et al.,* 2010) or biodegradable poly(D,L-lactide) (PLA) scaffolds (Kanczler *et al.,* 2008) with VEGF incorporated. Bone formation was measured using μCT and histology with both materials providing enhanced regeneration of the segmental bone defect compared to groups without HBMSCs and growth factors. The same research group employed diffusion chambers that were implanted intraperitoneally in MF-1 nu/nu mice for 10 weeks. The diffusion chambers afforded interaction with the *in vivo* milieu without ingrowth or response from the host cells and thus allowed evaluation of seeded HBMSCs on a biomimetic collagen scaffold and BMP-2-encapsulated (PLA) scaffolds (Yang *et al.,* 2003; Yang *et al.,* 2004a; Yang *et al.,* 2004b).

Subcutaneous models

The subcutaneous implant is a widely used model, usually, to assess the osteoconductive and osteoinductive properties of a material to develop ectopic bone. The disadvantage of the subcutaneous model is the perceived lack of clinical relevance and, typically, the absence of bone cells that can affect the potential bone formation capacity of a material. Nevertheless, the absence of local bone cells can also be advantageous and has been used for evaluation of HBMSCs that were seeded on solid hydroxyapatite and collagen scaffolds or in polysaccharide capsules (Dawson *et al.,* 2008; Pound *et al.,* 2006). To visualise the vascularisation of tissue-engineered constructs containing HBMSCs/bone allograft and PLA, a subcutaneous model was used in athymic mice. The constructs were impacted and implanted subcutaneously and after 4 weeks perfused *in vivo* with Microfil prior to μCT scanning (Bolland *et al.,* 2008). The subcutaneous implants provide a rapid screening model for "biocompatibility" and osteo-inductivity/-conductivity and have been used for bioactive materials carrying growth factors such as Bone morphogenetic proteins (BMPs) (Kisiel *et al.,* 2013). Several implants can be compared in the same individual giving paired data and hence the possibility to reduce the number of animals used while achieving valid information.

Critical sized cranial-defect models

Critical sized cranial defects (CSD) have been employed for bone material evaluation for over fifty years. As far back as 1957, Ray and Hollow described a cranial defect in rats that was treated with frozen intact bone, deproteinised bone or decalcified bone. Subsequently, Ko *et al*. used an 8 mm cranial defect in rats. However, the cranial defect model is technically challenging with risk to the dura mater as well as a risk of severe haemorrhage due to the presence of major blood vessels in the cranial bone (Ko *et al.,* 2008). Ray and Holloway circumvented these issues by creating smaller defects on the lateral side of the cranium to avoid the underlying blood vessels (Ray and Holloway, 1957). Given the smaller defect size (4 mm), two defects can be generated in the parietal bones lateral to the mid-sagittal suture in the cranium of rats and mice, allowing paired analysis and thus enhancing statistical power (Meinel *et al.,* 2005; Ventura *et al.,* 2014).

Femoral condyle-defect model

The use of large animal models has also been examined. Ueng *et al.* created bone cavities in the lateral femoral condyles of rabbits and implanted beads of alginate mixed with PKH 26-labelled rabbit mesenchymal stem cells and vancomycin (Ueng *et al.,* 2007). The beads of alginate

Table 2. *In vitro* assays with participating laboratories and numbers of separate studies *per in vitro* assay.

Assay	Parameter name
24 separate Alkaline Phosphatase Assays from 5 laboratories	Group 1
14 Biocompatibility Assays from 4 laboratories	Group 2
8 Calcium Deposition data sets from 2 laboratories	Group 3
20 Gene Expression data sets of early markers from 3 laboratories	Group 4
22 Gene Expression data sets of late markers from 3 laboratories	Group 5

demonstrated sustained release of vancomycin over 14 d and osteogenic differentiation of cultured mesenchymal stem cells (MSCs) in the alginate carrier matrix. The implanted MSCs contributed with newly formed bone *in vivo*.

Data analysis

The overall correlation between *in vitro* and *in vivo* outcomes across the entire dataset was characterised by sorting and then categorising the data according to *in vitro* score, before plotting the mean *in vivo* score obtained for each respective *in vitro* score category. Thus, for every biomaterial variable scoring 1, 2, 3 *etc. in vitro*, the corresponding *in vivo* scoring was listed and the mean plotted against each respective *in vitro* score 'category'. The same approach was adopted with the *in vivo* scores, which were again categorised to obtain the corresponding mean *in vitro* score for correlation.

As well as an overall assessment of *in vitro – in vivo* correlation, the *in vitro* outcomes were sorted into sub-groups representing individual classes of *in vitro* assay. The sub-groups were, i) Group 1; alkaline phosphatase activity, ii) Group 2; biocompatibility, iii) Group 3; calcium deposition iv) Group 4 and 5; gene expression of early markers and late markers respectively (Table 2). The *in vitro* outcome scores of the material were subsequently correlated with the *in vivo* outcome scores. Further, the *in vitro* assay groups were combined in pairs to investigate if a combination of *in vitro* assays provided a better prediction of *in vivo* outcome over single assays for correlation.

For the purpose of the remaining part of this paper, the following terms are used as defined below: Groups consist of *in vitro* assay (See Table 2), which were used as parameters to correlate the predictive outcome.

Null hypothesis: No correlation exists between *in vivo* and *in vitro* outcomes.

Hypothesis: Specific *in vitro* parameters or a combination of *in vitro* parameters can be used to predict material *in vivo* outcomes.

Correlation

Pearson's correlation was used to test for significant linear relationships between the *in vitro* and *in vivo* outcome ($n > 5$). Coefficients of determination (R^2) delineating the percentage of shared variance are presented and a p-value of < 0.05 was considered significant.

Sensitivity and specificity

The data for each *in vitro* assay groups was split into quartiles to distinguish false positives, false negatives, true positives and true negatives where a positive result was defined as one that scored > 2.5 and a negative result was defined as one that scored < 2.5 (Fig. 3**a**). Fisher's exact test was performed to measure the sensitivity and specificity of the various groups. A confidence interval of 95 % was used.

Results and Discussion

Despite considerable consensus in how assays were interpreted, the current analysis demonstrated no significant overall correlation between *in vitro* outcomes and *in vivo* outcomes.

The mean *in vitro* scores revealed a trend of covariance to *in vivo* scores with 58 %. The mean *in vivo* scores shared 51 % of variance when correlated to the *in vitro* categories (Fig. 1**a**). Analysis of the different *in vitro* techniques; biocompatibility, cell differentiation, gene expression of early markers, gene expression of late markers and calcium deposition demonstrated a covariance of less than 10 % (Fig. 1**b-f**). To determine if combinations of *in vitro* assays could predict the *in vivo* outcome, a selection was made that included materials evaluated by more than one *in vitro* assay. The medians of the combined *in vitro* scores were subsequently correlated to the *in vivo* outcomes (Fig. 2**a-f**). On analysis, less than 10 % covariance was observed except for the combination of Group 1; alkaline phosphatase expression with Group 2; biocompatibility assays, where a 95 % covariance was observed (Fig. 1**a**).

Our approach has its limitations. Any cross centre study, especially adopting a retrospective approach, is subject to the significant challenge of inevitable variation in protocols, reagents and cell source *etc*. between laboratories. Importantly, the main readout of the current study, the correlation between *in vitro* and *in vivo* outcomes will remain fairly robust against such variation as the point of comparison is not between the assays conducted across the various groups, but between the *in vitro* and *in vivo* stages of biomaterial testing undertaken in each individual laboratory. Perhaps more problematic is the inevitably limited scoring system required to tabulate the large variety and complexity of the assays conducted. For example, we reduce a complex readout such as osteogenic gene expression to a score of high and low expression of early and late markers. This is indeed a significant limitation

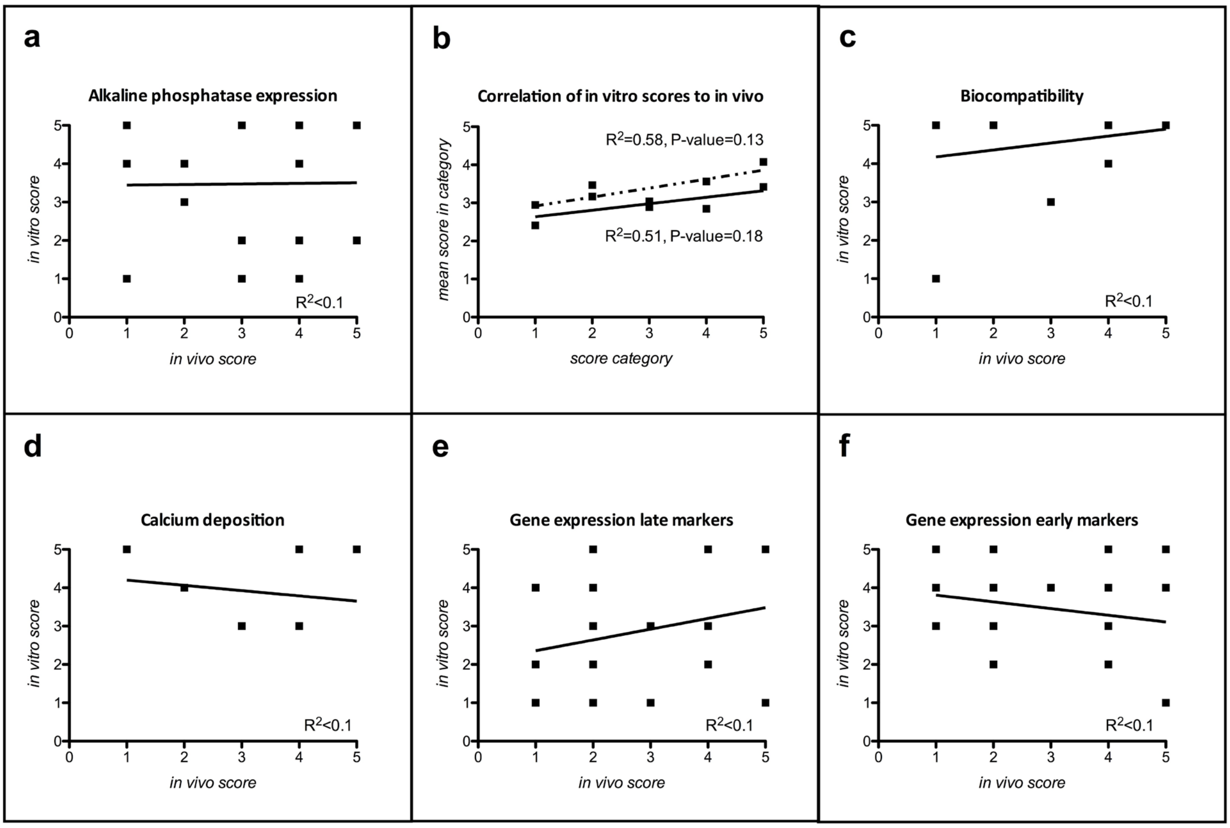

Fig. 1. Correlation of 47 *in vitro* studies and 36 *in vivo* studies testing 93 materials collected from 8 universities in a European multicentre study. **a**) The overall correlation between the *in vitro* and *in vivo* outcome was analysed by correlating the *in vitro* score category to the mean of the *in vivo* score in that category. Same method was used on the *in vivo* score category, where the mean of the *in vitro* assays was correlated to their corresponding *in vivo* score category. The mean *in vitro* scores had a covariance of 58 % to the *in vivo* score. The mean *in vivo* scores shared 51 % of variance with the *in vitro* categories. The data were sorted into subgroups of different *in vitro* assays. The subgroups were biocompatibility, alkaline phosphatase expression, calcium deposition, gene expression of early markers and gene expression of late markers. The *in vitro* outcome scores of the material were then correlated with the *in vivo* outcome scores. **b**) The covariance of the *in vitro* assay of Group 1; alkaline phosphatase expression to the scores of the same material *in vivo* correlating 24 separate studies (*n* = 24), **c**) The covariance between *in vivo* and *in vitro* of 12 separate *in vitro* assay of Group 2; biocompatibility (*n* = 12). **d**) The covariance between *in vivo* and *in vitro* of 8 separate *in vitro* assay of Group 3; calcium deposition (*n* = 8). **e**) The covariance of 18 separate *in vitro* assays of Group 4; early markers (*n* = 18) and **f**) Group 5; gene expression late markers (*n* = 18). All the subgroups were correlated to the *in vivo* outcome score.

of the current approach, but such broad-brush stroke assessments of performance reflect the conclusions often drawn from *in vitro* assays across the biomaterial literature. Such assessments thus serve as the hypotheses to be tested by our approach. Similarly, the qualitative assessment of each assay outcome was conducted by the groups that undertook the original research, raising possible concerns about subjective bias in the assessment of the results. This needs to be acknowledged as a clear limitation in the interpretation of the *in vitro* outcome itself and subsequent *in vivo* significance, but it is precisely such interpretations (we contend) of the success or failure of the biomaterial in the *in vitro* assays that currently form the basis for future *in vivo* testing and translation. Thus, this very limitation further serves the purpose of our study to test the decision making process underlying the biomaterial testing pipeline from *in vitro* to *in vivo* assessments and, it can be noted, justifies the contention of this review that such a lack of correlation is indeed surprising.

This surprisingly poor correlation between *in vitro* and *in vivo* assessments of biomaterials could be due to a number of further factors intrinsic to the assays themselves. These include cell choice, cell line, cell passage, and cell culture protocols, which while broadly similar between laboratories, can result in significantly different data outcomes. In a literature study, Bara *et al.* described the differences between naive MSCs and MSCs cultured *in vitro* for expansion. The authors reported a large effect of the isolation and culture parameters that gave opposing results in two clinical trials, using protocols with differences in density media, centrifugation steps and the combination of media and serum. Interestingly, the seeding of whole bone marrow had a positive effect on Colony forming unit (CFU) efficiency and telomere length of the

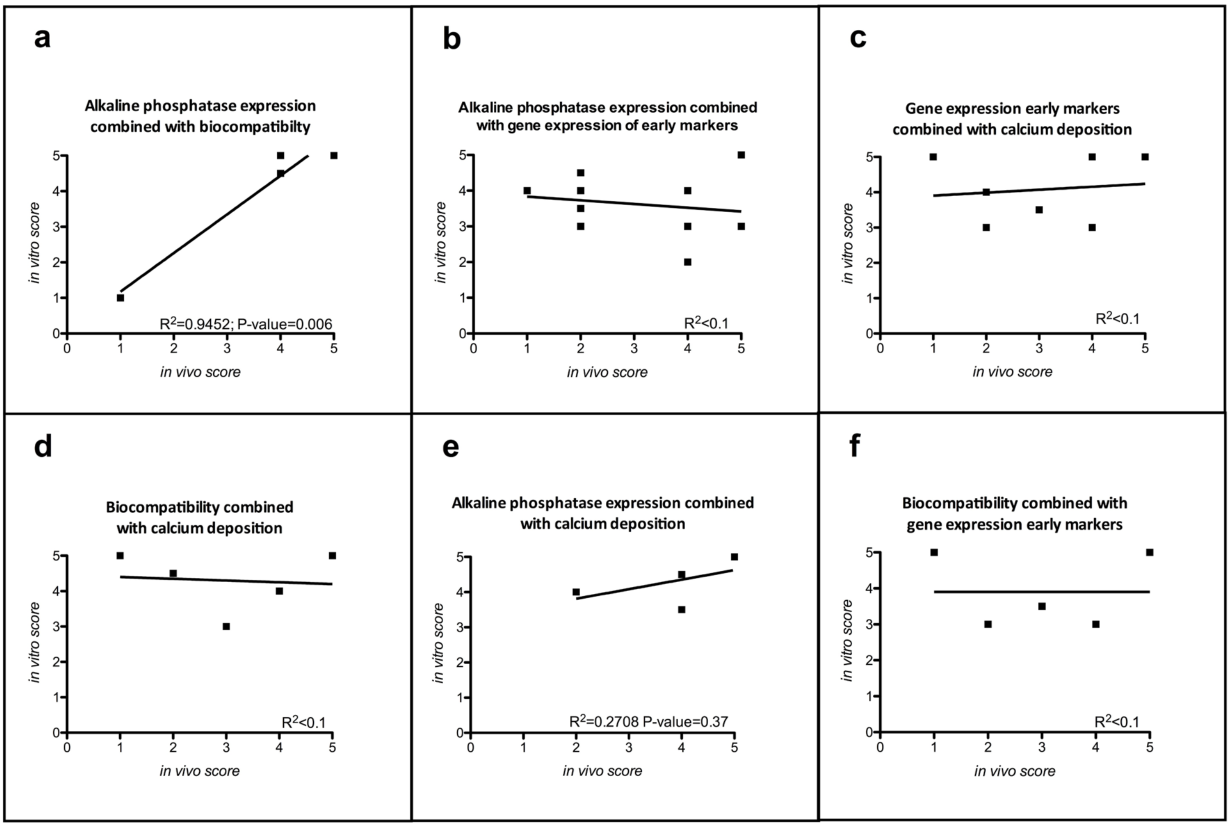

Fig. 2. The groups were combined in pairs to investigate if the combination could predict the *in vivo* outcome. **a**) Combined *in vitro* score of Group 1; alkaline phosphatase and Group 2; biocompatibility (n = 5); here a significant covariance of 94.5 % was seen with a p-value of 0.0055. **b**) Combined *in vitro* scores of Group 1; alkaline phosphatase and Group 3; calcium deposition (n = 5). **c**) Combined *in vitro* score of Group 1; alkaline phosphatase expression and Group 4; gene expression of early markers (n = 11). **d**) Combined *in vitro* score of Group 2; biocompatibility and Group 3; calcium deposition (n = 5). **e**) Combined *in vitro* score of Group 2; biocompatibility and Group 4; gene expression of early markers (n = 5). **f**) Group 4; gene expression early markers combined with Group 3; calcium deposition (n = 7). All group combinations except for Group 1/Group 2 and Group 1/Group 3 shared covariance of less than 10 %.

MSCs. Furthermore, a lower seeding density enhanced proliferation while the application of autologous serum had a positive effect on MSC stability and maintenance of the naive state. The sensitivity of these parameters and their significance for clinical outcome thus underlines the importance of standardisation of protocols in order to dissect differences in the response to biomaterials against inherent differences in MSC phenotype due to various expansion protocols (Bara *et al.,* 2014).

A further potential confounding factor, particularly for *in vitro* assays using primary cells, is the known issue of donor variation (Georgi *et al.,* 2015). To try and overcome donor cell variation when assaying biomaterials for cell responses, several studies have pooled cells from multiple donors (Stoddart *et al.,* 2012). This approach has the additional benefit of allowing larger cell numbers to be obtained without extensive cell expansion. Stoddart *et al.*, however, questioned the usefulness of this *in vitro* strategy for predicting *in vivo* outcomes. Arguably, donor cell variation *in vitro* mimics the host cell variation encountered *in vivo* – a highly relevant variable for predicting *in vivo* response which is obscured by the artificial donor/host situation achieved through pooling multiple cell sources. Though experimentally more intensive, biomaterials should ideally be tested using single donor cells repeated across several donors in order to quantify the robustness of the biomaterial strategy against donor/host variation.

An alternative approach to overcoming the confounding influence of donor cell variation is the evaluation of materials using cell lines. While primary cells would appear likely to better approximate host responses, including, as above, host variation, such variation in the absence of large multi donor analyses, makes comparison of cell-biomaterial responses between studies difficult. A key advantage afforded by the implementation of cell lines is the homogeneity of the cell population allowing, at least at the early stages of the biomaterial testing pipeline, relatively straightforward comparison of results across studies and laboratories. Furthermore, cell lines can often allow differentiation stage-specific responses to biomaterials. For example, one study comparing the ability of various cell lines to mimic the response of mature osteoblasts found the MC3T3-E1 cell line to predict primary osteoblast responses *in vitro* with high fidelity (Czekanska *et al.,* 2012).

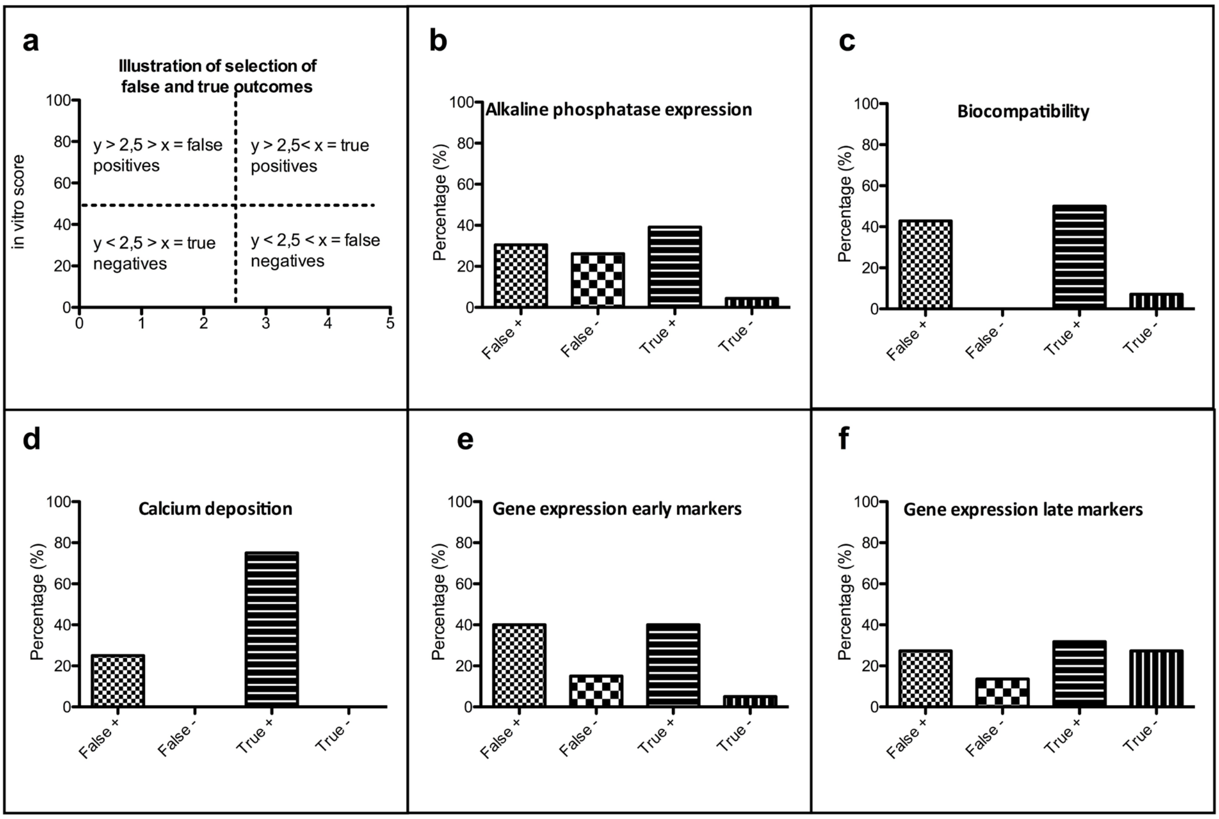

Fig. 3. Each study of the groups was split into quartiles to sort them into percentage of false positives, false negatives, true positives and true negatives where a positive result was one that scored > 2.5 and a negative result one that scored < 2.5. **a**) Illustration of the quartiles selected with score 2.5 as a cut off for testing sensitivity *versus* specificity. **b**) Group 1; alkaline phosphatase expression (*n* = 24) detected true positives to a marginally greater extent than false positives. **c**) Group 2; biocompatibility (*n* = 12) detected true positives to a slightly greater extent than false positives (50 %). **d**) Group 3; calcium deposition (*n* = 8) detected a majority of true positives 75 %. Both **e**) Group 4; gene expression early markers (*n* = 18) and **f**) Group 5; gene expression late markers (*n* = 18) showed a random detection with all four outcomes. All the groups were tested with Fisher's exact test to further measure the sensitivity and specificity on the various groups; all *p* values obtained were of statistical non-significance.

Nevertheless, even accounting for the variations in experimental protocols and cell sources for each assay method, almost all *in vitro* assays demonstrated the ability to correctly identify positive outcomes (*i.e.* true positives) to a slightly greater extent than false positives (*i.e.* incorrectly predict a positive outcome). The statistical power of the *in vitro* assay groups was variable, thus affecting the ability to interpret the results accurately. However, increasing the number of studies *per in vitro* assay group to increase the power of each predictive model may not necessarily improve the specificity of the model as evidenced by Fig. 1**a**, **e** and **f**. Fisher's exact test was performed to further measure the sensitivity and specificity on the various groups; all *p*-values obtained were of statistical non-significance and thus inconclusive (Fig. 3-4). One approach to improve the predictability would be to develop a system of metrics based on the combined score achieved through a series of *in vitro* assays. Thus, for example, the combined *in vitro* tests of biocompatibility and alkaline phosphatase expression had a strong positive correlation, indicating a potential benefit of *in vitro* assay combinations. The combination of groups demonstrated the ability to detect true positives to a greater extent than false positives in all groups.

It is self-evident that a large selection of *in vivo* models are used to assess the performance of biomaterials tailored for bone. Each *in vivo* model is often slightly modified for the purpose of the particular study and to suit the properties of the biomaterial tested. Thus, the large number of *in vivo* models available can make direct comparison difficult. Standardised and validated defects with defined parameters of *in vivo* outcome would facilitate direct comparison and minimise the number of animals used (Kilkenny *et al.*, 2010; Reichert *et al.*, 2009; van Griensven, 2015). It is also necessary to appreciate that not only do the *in vivo* models differ, but also definitions of successful bone regeneration differ between each study.

We have highlighted the surprisingly poor correlation between *in vitro* success or failure and *in vivo* success or failure in biomaterial testing for bone repair. It should also be noted that this challenge may be relatively modest compared with the subsequent challenges of achieving successful translation from small to large animal models and from preclinical studies into clinical trials (van

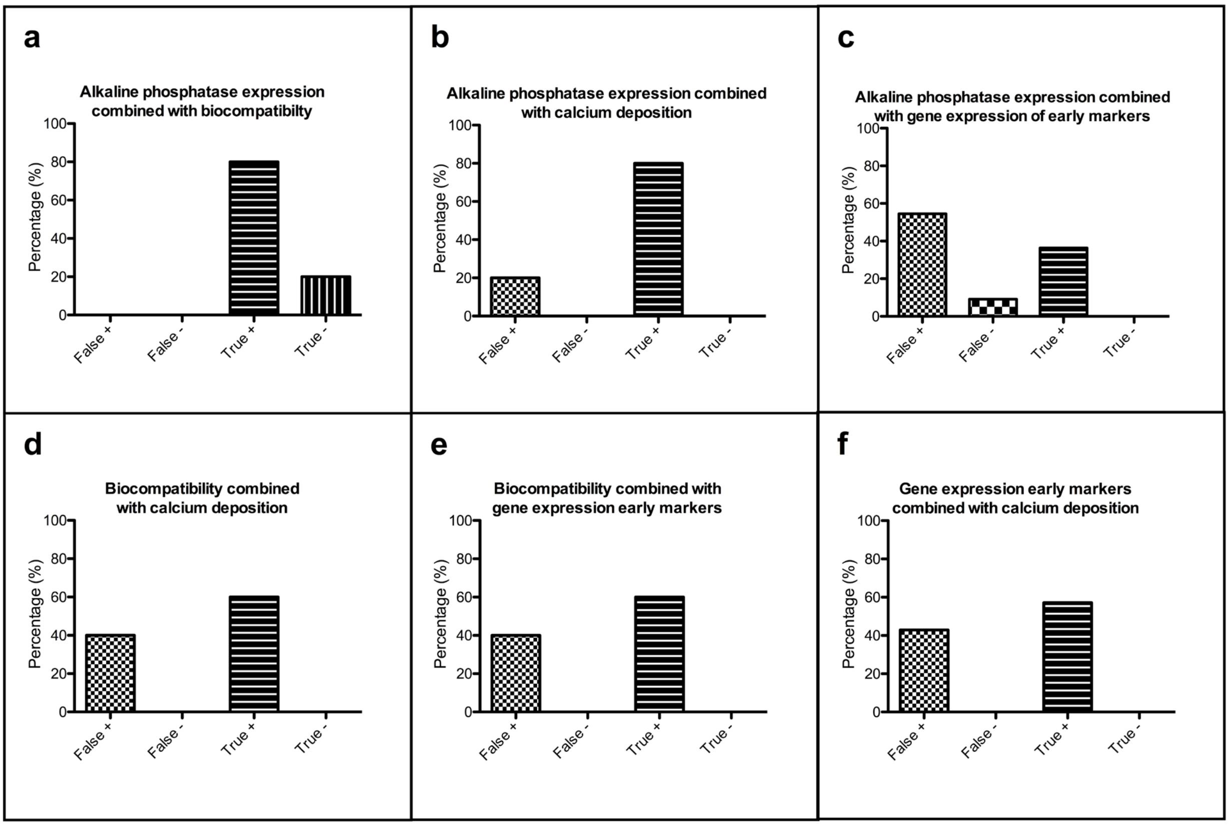

Fig. 4. The groups were combined in pairs to investigate if a combination could predict the *in vivo* outcome. Each study of the groups was split into quartiles to sort them into percentage of false positives, false negatives, true positives and true negatives where a positive result was one that scored > 2.5 and a negative result one that scored < 2.5. **a**) Group 1; alkaline phosphatase and Group 2; biocompatibility (*n* = 5) detected true positives and true negatives, **b**) Group 1; alkaline phosphatase and Group 3; calcium deposition (*n* = 5) detected 20 % false positives and 80 % true positives. **c**) Group 1; alkaline phosphatase expression and Group 4; gene expression of early markers (*n* = 11) detected a large amount of false positives 55 % compared to 36 % true positives and a small amount of false negatives was also detected. Both **d**) Group 2; biocompatibility and Group 3; calcium deposition (*n* = 5) and **e**) Group 2; biocompatibility and Group 4; gene expression of early markers, detected a majority of true positives 60 %. **f**) Group 4; gene expression early markers combined with Group 3; calcium deposition (*n* = 7) had a slight majority of 57 % of true positives. The combination of groups demonstrated the ability to detect true positives to a greater extent than false positives in all groups. Fisher's exact test was performed to further measure the sensitivity and specificity on the various groups; all *p* values obtained were of statistical non-significance and thus inconclusive.

Griensven, 2015). Critically, at each stage, increased cost as well as surgical and physiological complexity, challenge reproducibility as well as predictive power (Reichert *et al.,* 2009). Hackam and Redelmeier explored the translation of research from animals to humans and concluded that only about a third of highly cited preclinical articles were later translated to human trials (Hackam and Redelmeier, 2006). One promising approach to addressing these inherent limitations of large animal models has been the development of humanised animal models. Humanised animal models may, finally, serve to improve the predictive power of pre-clinical assessments (Holzapfel *et al.,* 2015).

Conclusion

While we have here noted the weakness of several widely used *in vitro* measures of biomaterial success, it is important to emphasise the vital importance of the application of *in vitro* assays for biomaterial testing for bone regenerative medicine. The growing number of potential biomaterials for application in bone regeneration strategies warrants increased research investment in the development of early stage assays predictive of *in vivo* success or failure in order to streamline the biomaterial testing pipeline towards the goal of clinical translation. The current inadequacies of our *in vitro* assays highlighted here, will we hope further underline the urgency of this research imperative and, critically, stimulate the development of novel approaches to biomaterial testing and appropriately characterised pipelines with measurable sensitivity and specificity for translatable *in vivo* outcomes. In this way, we hope the rather long-awaited therapeutic promise of bone regenerative medicine will begin to be realised.

Acknowledgements

This work was supported by a European Community Seventh Framework Programme Grant, BioDesign (262948). We acknowledge the provision of data from members of the EU BioDesign consortium, without whom this work would not have been possible. We thank Wictor Billström for valuable technical assistance in data sorting as well as Daniel Tang and Stefanie Inglis for constructive comments on the manuscript.

References

Bara JJ, Richards RG, Alini M, Stoddart MJ (2014) Concise review: Bone marrow-derived mesenchymal stem cells change phenotype following *in vitro* culture: implications for basic research and the clinic. Stem Cells **32**: 1713-1723.

Bartal BA I, Decety J, Mason P (2011) Empathy and pro-social behavior in rats. Science **334**: 1427-1430.

Bering J (2012)The Rat that laughed. In: Scientific American, Nature Publishing Group, pp 74-77.

Billstrom GH, Blom AW, Larsson S, Beswick AD (2013) Application of scaffolds for bone regeneration strategies: current trends and future directions. Injury **44 Suppl 1**: S28-S33.

Bolland BJ, Kanczler JM, Dunlop DG, Oreffo RO (2008) Development of *in vivo* μCT evaluation of neovascularisation in tissue engineered bone constructs. Bone **43**: 195-202.

Bonewald LF (2006) Mechanosensation and Transduction in Osteocytes. Bonekey Osteovision **3**: 7-15.

Bonewald LF (2011) The amazing osteocyte. J Bone Miner Res **26**: 229-238.

Bruderer M, Richards RG, Alini M, Stoddart MJ (2014) Role and regulation of RUNX2 in osteogenesis. Eur Cell Mater **28**: 269-286.

Czekanska EM, Stoddart MJ, Richards RG, Hayes JS (2012) In search of an osteoblast cell model for *in vitro* research. Eur Cell Mater **24**: 1-17.

Dawson AB (1926) A note on the staining of the skeleton of cleared specimens with alizarin red S. Biotechnic & Histochemistry **1**: 123-124.

Dawson JI, Wahl DA, Lanham SA, Kanczler JM, Czernuszka JT, Oreffo RO (2008) Development of specific collagen scaffolds to support the osteogenic and chondrogenic differentiation of human bone marrow stromal cells. Biomaterials **29**: 3105-3116.

Einhorn TA, Lane JM, Burstein AH, Kopman CR, Vigorita VJ (1984) The healing of segmental bone defects induced by demineralized bone matrix. A radiographic and biomechanical study. J Bone Joint Surg Am **66**: 274-279.

F1983-14 A Practice for Assessment of Selected Tissue Effects of Absorbable Biomaterials for Implant Applications. ASTM International.

F2721-09r14 A guide for pre-clinical *in vivo* evaluation in critical size segmental bone defects. ASTM International.

Frost HM (1994) Wolff's Law and bone's structural adaptations to mechanical usage: an overview for clinicians. The Angle Orthod **64**: 175-188.

Garcia-Gareta E, Coathup MJ, Blunn GW (2015) Osteoinduction of bone grafting materials for bone repair and regeneration. Bone **81**: 112-121.

Georgi N, Taipaleenmaki H, Raiss CC, Groen N, Portalska KJ, van Blitterswijk C, de Boer J, Post JN, van Wijnen AJ, Karperien M (2015) MicroRNA levels as prognostic markers for the differentiation potential of human mesenchymal stromal cell donors. Stem Cells Dev **24**: 1946-1955.

Giannoudis PV, Einhorn TA, Marsh D (2007) Fracture healing: the diamond concept. Injury **38 Suppl 4**: S3-S6.

Giannoudis PV, Einhorn TA, Schmidmaier G, Marsh D (2008) The diamond concept--open questions. Injury **39 Suppl 2**: S5-S8.

Gibbs DM, Black CR, Dawson JI, Oreffo RO (2014) A review of hydrogel use in fracture healing and bone regeneration. J Tissue Eng Regen Med **10**: 187-198.

Goodwin CJ, Holt SJ, Downes S, Marshall NJ (1995) Microculture tetrazolium assays: a comparison between two new tetrazolium salts, XTT and MTS. J Immunol Methods **179**: 95-103.

Gregory CA, Gunn WG, Peister A, Prockop DJ (2004) An Alizarin red-based assay of mineralization by adherent cells in culture: comparison with cetylpyridinium chloride extraction. Anal Biochem **329**: 77-84.

Hackam DG, Redelmeier DA (2006) Translation of research evidence from animals to humans. JAMA **296**: 1731-1732.

Hollinger JO, Kleinschmidt JC (1990) The critical size defect as an experimental model to test bone repair materials. J Craniofac Surg **1**: 60-68.

Holzapfel BM, Hutmacher DW, Nowlan B, Barbier V, Thibaudeau L, Theodoropoulos C, Hooper JD, Loessner D, Clements JA, Russell PJ, Pettit AR, Winkler IG, Levesque JP (2015) Tissue engineered humanized bone supports human hematopoiesis *in vivo*. Biomaterials **61**: 103-114.

Kaipel M, Schutzenberger S, Schultz A, Ferguson J, Slezak P, Morton TJ, Van Griensven M, Redl H (2012) BMP-2 but not VEGF or PDGF in fibrin matrix supports bone healing in a delayed-union rat model. J Orthop Res **30**: 1563-1569.

Kanczler JM, Ginty PJ, Barry JJ, Clarke NM, Howdle SM, Shakesheff KM, Oreffo RO (2008) The effect of mesenchymal populations and vascular endothelial growth factor delivered from biodegradable polymer scaffolds on bone formation. Biomaterials **29**: 1892-1900.

Kanczler JM, Ginty PJ, White L, Clarke NM, Howdle SM, Shakesheff KM, Oreffo RO (2010) The effect of the delivery of vascular endothelial growth factor and bone morphogenic protein-2 to osteoprogenitor cell populations on bone formation. Biomaterials **31**: 1242-1250.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol **8**: e1000412.

Kirker-Head C, Karageorgiou V, Hofmann S, Fajardo R, Betz O, Merkle HP, Hilbe M, von Rechenberg B, McCool J, Abrahamsen L, Nazarian A, Cory E, Curtis M, Kaplan D, Meinel L (2007) BMP-silk composite matrices heal critically sized femoral defects. Bone **41**: 247-255.

Kisiel M, Martino MM, Ventura M, Hubbell JA, Hilborn J, Ossipov DA (2013) Improving the osteogenic potential of BMP-2 with hyaluronic acid hydrogel modified with integrin-specific fibronectin fragment. Biomaterials **34**: 704-712.

Knight A (2008) Non-animal methodologies within biomedical research and toxicity testing. Altex **25**: 213-231.

Ko EK, Jeong SI, Rim NG, Lee YM, Shin H, Lee BK (2008) *In vitro* osteogenic differentiation of human mesenchymal stem cells and *in vivo* bone formation in composite nanofiber meshes. Tissue Eng Part A **14**: 2105-2119.

Langley G, Evans T, Holgate ST, Jones A (2007) Replacing animal experiments: choices, chances and challenges. Bioessays **29**: 918-926.

Marsell R, Einhorn TA (2009) The role of endogenous bone morphogenetic proteins in normal skeletal repair. Injury **40 Suppl 3**: S4-S7.

Marsell R, Einhorn TA (2011) The biology of fracture healing. Injury **42**: 551-555.

McComb RB, Bowers GN, Posen S (1979) Alkaline phosphatase. New York: Plenum Press **XVI:** 986 p.

Meinel L, Betz O, Fajardo R, Hofmann S, Nazarian A, Cory E, Hilbe M, McCool J, Langer R, Vunjak-Novakovic G, Merkle HP, Rechenberg B, Kaplan DL, Kirker-Head C (2006) Silk based biomaterials to heal critical sized femur defects. Bone **39**: 922-931.

Meinel L, Fajardo R, Hofmann S, Langer R, Chen J, Snyder B, Vunjak-Novakovic G, Kaplan D (2005) Silk implants for the healing of critical size bone defects. Bone **37**: 688-698.

Millan JL (2013) The role of phosphatases in the initiation of skeletal mineralization. Calcif Tissue Int **93**: 299-306.

Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D (2011) The cost of drug development: a systematic review. Health Policy **100**: 4-17.

Müller U (2007) *In vitro* biocompatibility testing of biomaterials and medical devices. Med Device Technol **19**: 32-34.

Panksepp J, Burgdorf J (2003) "Laughing" rats and the evolutionary antecedents of human joy? Physiol Behav **79**: 533-547.

Pound JC, Green DW, Chaudhuri JB, Mann S, Roach HI, Oreffo RO (2006) Strategies to promote chondrogenesis and osteogenesis from human bone marrow cells and articular chondrocytes encapsulated in polysaccharide templates. Tissue Eng **12**: 2789-2799.

Ray RD, Holloway JA (1957) Bone implants; preliminary report of an experimental study. J Bone Joint Surg Am **39-A**: 1119-1128.

Reichert JC, Saifzadeh S, Wullschleger ME, Epari DR, Schutz MA, Duda GN, Schell H, van Griensven M, Redl H, Hutmacher DW (2009) The challenge of establishing preclinical models for segmental bone defect research. Biomaterials **30**: 2149-2163.

Russell WMS, Burch RL (1959) The principles of humane experimental technique. London: Methuen 238 p.

Schutzenberger S, Schultz A, Hausner T, Hopf R, Zanoni G, Morton T, Kropik K, van Griensven M, Redl H (2012) The optimal carrier for BMP-2: a comparison of collagen *versus* fibrin matrix. Arch Orthop Trauma Surg **132**: 1363-1370.

Stoddart MJ, Richards RG, Alini M (2012) *In vitro* experiments with primary mammalian cells: to pool or not to pool? Eur Cell Mater **24**: i-ii.

Ueng SW, Lee MS, Lin SS, Chan EC, Liu SJ (2007) Development of a biodegradable alginate carrier system for antibiotics and bone cells. J Orthop Res **25**: 62-72.

van Griensven M (2015) Preclinical testing of drug delivery systems to bone. Adv Drug Deliv Rev **94**: 151-164.

van Wijnen AJ, Stein GS, Gergen JP, Groner Y, Hiebert SW, Ito Y, Liu P, Neil JC, Ohki M, Speck N (2004) Nomenclature for Runt-related (RUNX) proteins. Oncogene **23**: 4209-4210.

Ventura M, Boerman OC, Franssen GM, Bronkhorst E, Jansen JA, Frank Walboomers XF (2014) Monitoring the biological effect of BMP-2 release on bone healing by PET/CT. J Control Release **183:** 138-144.

Vodovotz Y, Chow CC, Bartels J, Lagoa C, Prince JM, Levy RM, Kumar R, Day J, Rubin J, Constantine G, Billiar TR, Fink MP, Clermont G (2006) *In silico* models of acute inflammation in animals. Shock **26**: 235-244.

Yang X, Tare RS, Partridge KA, Roach HI, Clarke NM, Howdle SM, Shakesheff KM, Oreffo RO (2003) Induction of human osteoprogenitor chemotaxis, proliferation, differentiation, and bone formation by osteoblast stimulating factor-1/pleiotrophin: osteoconductive biomimetic scaffolds for tissue engineering. J Bone Miner Res **18**: 47-57.

Yang XB, Bhatnagar RS, Li S, Oreffo RO (2004a) Biomimetic collagen scaffolds for human bone cell growth and differentiation. Tissue Eng **10**: 1148-1159.

Yang XB, Whitaker MJ, Sebald W, Clarke N, Howdle SM, Shakesheff KM, Oreffo RO (2004b) Human osteoprogenitor bone formation using encapsulated bone morphogenetic protein 2 in porous polymer scaffolds. Tissue Eng **10**: 1037-1045.

Editor's Notes: All questions/comments by the reviewers were answered by text changes. There is hence no Discussion with Reviewers section.

Scientific Editor in charge of the paper: Joost de Bruijn.