Regulatory Toxicology and Pharmacology 73 (2015) 367–377

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Establishing best practise in the application of expert review of mutagenicity under ICH M7

Chris Barber [a, *], Alexander Amberg [b], Laura Custer [c], Krista L. Dobo [d], Susanne Glowienke [e], Jacky Van Gompel [f], Steve Gutsell [g], Jim Harvey [h], Masamitsu Honma [i], Michelle O. Kenyon [d], Naomi Kruhlak [j], Wolfgang Muster [k], Lidiya Stavitskaya [j], Andrew Teasdale [l], Jonathan Vessey [a], Joerg Wichard [m]

[a] Lhasa Limited, Leeds, UK
[b] Sanofi-Aventis Deutschland GmbH, DSAR Preclinical Safety, Frankfurt, Germany
[c] Bristol-Myers Squibb, Drug Safety Evaluation, New Brunswick, USA
[d] Pfizer, Drug Safety Research and Development, Groton, CT, USA
[e] Novartis Institutes for Biomedical Research, Department of Preclinical Safety, Basel, Switzerland
[f] Janssen, Drug Safety Sciences, Beerse, Belgium
[g] Unilever, Safety and Environmental Assurance Centre, Colworth, Beds, UK
[h] GlaxoSmithkline, Computational Toxicology, Ware, Herts, UK
[i] National Institute of Health Sciences, Tokyo, Japan
[j] FDA Center for Drug Evaluation and Research, Silver Spring, MD, USA
[k] F. Hoffmann-La Roche Ltd., Pharma Research and Early Development, Basel, Switzerland
[l] AstraZeneca, Macclesfield, Cheshire, UK
[m] Bayer, HealthCare, Genetic Toxicology, Berlin, Germany

ARTICLE INFO

Article history:
Received 19 July 2015
Received in revised form
21 July 2015
Accepted 22 July 2015
Available online 4 August 2015

Keywords:
ICH M7
Ames
Genotoxicity
Mutagenicity
Expert rule-based
Statistical
In silico

ABSTRACT

The ICH M7 guidelines for the assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals allows for the consideration of *in silico* predictions in place of *in vitro* studies. This represents a significant advance in the acceptance of (Q)SAR models and has resulted from positive interactions between modellers, regulatory agencies and industry with a shared purpose of developing effective processes to minimise risk. This paper discusses key scientific principles that should be applied when evaluating *in silico* predictions with a focus on accuracy and scientific rigour that will support a consistent and practical route to regulatory submission.

1. Introduction

The ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) is an international group comprised of regulatory authorities and industry across Europe, Japan and the US which has taken a predominant position in establishing guidelines to support the development and registration of safe and effective medicines. Since 1990, this body has established almost 50 guidelines covering many processes from manufacturing quality to clinical trial design and drug safety. One of these (M7) focusses upon the assessment of potentially DNA-reactive (mutagenic) impurities and is significant in that, it recognises the potential to use *in silico* predictions in lieu of *in vitro* studies. This substitution is not necessarily a simple process because of the inherent uncertainty that exists with any *in silico* prediction. However, the risk of not identifying a potentially DNA-reactive impurity that is subsequently exposed to humans can

* Corresponding author. Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS, UK.
E-mail address: chris.barber@lhasalimited.org (C. Barber).

http://dx.doi.org/10.1016/j.yrtph.2015.07.018
0273-2300/

be minimised through the appropriate use of well-constructed models and, crucially, through the use of expert analysis. Whilst this principle is clearly defined in the guidelines, the practical implementation of a process that can ensure consistent, safe and accurate predictions requires some consideration. This paper identifies key questions and approaches that an expert could apply in order to maximise sensitivity in the identification of DNA-reactive materials without necessarily applying an overly cautious approach that reduces specificity and overall accuracy to a point where *in silico* predictions offer little additional value.

The intention of the M7 guideline is to describe how best to identify and control the exposure to (potential) pharmaceutical agent impurities that could cause cancer through direct reaction with DNA. Such impurities may be introduced during the preparation of the active pharmaceutical ingredient (API), during formulation of the final drug product, or through degradation. DNA reactivity can be effectively tested using the *in vitro* bacterial reverse mutation assay (often somewhat loosely referred to as 'the Ames assay'), but this study may be impractical or unnecessary given our knowledge of the endpoint and weight of data already obtained. (Q)SAR is recommended in the ICH M7 guidelines as a high-throughput, state-of-the-art alternative for assessing the mutagenic potential of such impurities. This paper focusses not on the identification of such potential or observed impurities, but on the subsequent (Q)SAR analysis of their likelihood of being DNA-reactive.

The guideline describes the need for two predictive systems; one expert rule-based and the second statistical-based. The application of two systems that use different methods is predicated on the assumption that their predictions will be complementary and that greater sensitivity in detecting potential mutagens will be achieved if they are applied in a manner where a positive prediction from either method leads to a positive conclusion. Applying the models in this manner typically results in a decrease in specificity and overall accuracy; however, some of this decrease can be mitigated through the application of expert knowledge. Indeed, the guidelines make a specific provision for the application of expert knowledge to support or overturn a (Q)SAR prediction, effectively allowing a positive or negative *in silico* conclusion to be challenged through the rational consideration of additional information. While expert assessment has been successfully applied and reported (Dobo et al., 2012; Sutter et al., 2013), the definition of what constitutes expert analysis and how it may be undertaken has until now, been left largely open (Powley, 2015; Greene et al., 2015). Here we attempt to define a framework through which this process may be more clearly understood, although it should be emphasised that we do not anticipate this to be exhaustive or that when applied, experts will necessarily come to the same conclusions.

2. Comparison between expert rule-based and statistical systems

The distinction between expert rule-based and statistical systems is not necessarily clear-cut; experts may use statistical models to support their knowledge extraction during the development of an expert rule-based system, and the building of statistical systems undoubtedly benefits from oversight by an expert.

An expert rule-based system is comprised of rules written by humans which allows for the injection of knowledge in addition to known (in)activity of compounds. These include the biological mode(s) of action or expected metabolic transformations together with a chemical understanding of reactivity. This facilitates the provision of a detailed and clearly reasoned explanation for an alert along with references and mechanistic information – all of which can offer significant benefit during expert analysis. The injection of additional knowledge also allows the model builder to extrapolate from that which is known from the training set, for example to hitherto unseen functionality based upon an understanding of chemical reactivity, or through the incorporation of knowledge from proprietary data that can remain transparent and interpretable without revealing confidential structures. This can help an expert rule-based model maintain strong performance against chemical space more dissimilar to the training set than a statistical system can often achieve. Such an approach can also allow more complex endpoints to be modelled – for example if there are multiple modes of actions within a single dataset, then a simple global statistical model may struggle to correctly learn significant trends.[1] Some more sophisticated expert rule-based systems apply layers of reasoning which enables the model to continue to work in the presence of contradictory information – something which is important when modelling biological data since reproducibility is often not complete.

In contrast, a statistical system must rely upon machine-learning to determine the importance of a descriptor or combination of descriptors subsequently used to predict activity. For it to show a high degree of transparency, some expert knowledge almost always influences the model, most commonly through the choice of descriptors from which the model can learn. This can range from defining a list of descriptors to be used during model building (e.g. a predefined fragment library) or by defining rules by which descriptors should be created from the training set (e.g. fragmentation rules appropriate to the endpoint). Each approach offers benefits and challenges, but the result is still a statistical model provided that the decision as to the final selection of descriptors and impact that they make upon a prediction is learnt by the model through application of the training data. Statistical models tend to work best when predicting compounds of similar structure to those in the training set, and this has given rise to a number of approaches to defining applicability domains. Typically this is a set of criteria implemented by the modeller below which the predicted accuracy is considered insufficient for reliable use. However, this measure should be treated with caution, since even if the approach is explicit and the robustness demonstrated (sadly both are often missing), the modeller cannot know the use to which the model will be put and hence the acceptable level of accuracy that is desired. Models that apply the same descriptors for both predictions and determining the applicability domain and focus upon human interpretable properties that have demonstrable relevance to the endpoint are more likely to be able to provide transparent, relevant and unambiguous estimates of expected accuracy than those statistically tuned to a certain level of performance against a particular test set (Sahigara et al., 2012; Netzeva et al., 2005).

3. Ensuring *in silico* models can support expert analysis

Many models of mutagenicity have been developed using a range of methodologies. All have benefited from a large dataset of *in vitro* data tested under well established and mostly consistent conditions. This, together with a clear understanding of the mechanisms of action, has enabled models to make predictions which can be trusted provided the supporting evidence and rationale is readily accessible. Whilst some models may be used to illustrate points, this paper does not endorse any specific system

[1] Some statistical approaches can at least partly address this, for example by clustering compounds before constructing local models and thereby identify significant trends that would be swamped if the dataset was considered as a whole(Hanser et al., 2014).