Guidelines for rare disease model development and review

		Document Record ID Key	
Work stream			
Programme Director	Tom Fowler	**Status**	Final
Document Owner	Richard Scott	**Version**	5.1
Document Author	Andrew Devereau, Damian Smedley, Richard Scott	**Version Date**	9/5/16

Guidelines for rare disease model development and review

Summary

When developing new disease-specific phenotype models for rare diseases, or reviewing existing models:

1. Aim for between 20-40 terms
2. Ensure the core phenotype is captured
3. Aim to avoid very common terms or very rare terms
4. Avoid redundancy or unnecessary duplication
5. Aim for consistency of approach between disease models

Overall, the aim is to achieve a balance between too much detail and too little.

1.1 Background

The disease-specific data models for rare diseases aim to ensure that phenotype data collected for participants is detailed, specific and consistent. The HPO data models appear as disease-specific questionnaires, requiring the clinician to indicate whether a series of phenotypes represented by the HPO terms are present or absent, or their presence is unknown. They can also add modifiers to the terms which are present, e.g. to indicate severity, laterality or pace of progression. Additional HPO terms that are not listed in the questionnaire can be selected from the whole HPO ontology.

This questionnaire-based approach has been adopted to enable detailed data collection amongst expert and less disease-expert recruiters. It will be used along with the participant's genome for their diagnosis, and also to add essential clinical detail to the anonymised data set being developed by Genomics England for research.

To date we have produced >160 disease models and continue to add more. Feedback from users, regular reviews of the existing models and development of quality control measures have allowed us to develop this set of guidelines for model development and revision. These are guidelines rather than strict rules. Overall, the aim is to achieve a balance between too much detail and too little.

1.2 New model development

When a new disease is accepted for inclusion into the programme, we work with the disease proposer and a small disease expert team drawn from the GMC and GeCIP communities to identify a starting point for the model. This can be a published review of the disease, an existing CRF or database design or a clinical proforma. We translate this to HPO terms and clinical tests types and agree a draft model with the expert team for circulation to

all stakeholders and finalisation. Any terms that are required which are not in the HPO ontology can be flagged for future revision of the ontology.

1.3 Model review

Review of our full catalogue of rare disease HPO models is taking place during April 2016. This will identify any necessary revision of existing models. As with new model development, this will be conducted with small disease expert teams drawn from the GMC and GeCIP communities, supported by members of the Science and Bioinformatics teams at Genomics England and with input from the Rare Disease Clinical Data Working Group. Revised models will be circulated to all stakeholders for comment before adoption.

1.4 Model development and revision principles

The principles set out below will be used to guide the process of model development and revision.

1. Aim for between 20-40 terms

   Models with very few terms can lack the specificity needed to describe a disease in detail. Equally, those with very many terms are lengthy and difficult to complete. We are therefore aiming for 20-40 terms: models of this size achieve a good specificity and are practical to complete. Our analysis of our existing models using the Monarch phenotype sufficiency score* indicates that models of this size achieve good sufficiency, comparable with the larger models. However this is a guideline: the terms included in the model must be clinically relevant and necessary to describe the disease, and models outside the 20-40 terms range may be valid.

   Achieving this: the Genomics England team will identify models that are outside the 20-40 size range. Those substantially outside the target range will be prioritised for revision.

   **The Monarch sufficiency score tests that a disease phenotype description has 'necessary and sufficient information characteristics required to identify disease similarity based on phenotypes alone' (http://phenoday2014.bio-lark.org/pdf/6.pdf). We are developing a similar approach to guide model development using the Phenomizer algorithm (http://compbio.charite.de/phenomizer/).*

2. Ensure the core phenotype is captured

   In many cases there are one or more directly relevant HPO terms indicated by the disorder which should be included in the model. For example, the disease 'Hypertrophic cardiomyopathy' matches the HPO term of the same name. Inclusion of such terms is vital to allowing HPO-led analysis.

   When a participant is recruited it is important that data are captured for these core or disease defining terms. We are therefore tagging such terms during model review and new model development.

Achieving this: the Genomics England team will suggest suitable terms when developing new models and have reviewed all currently models for relevant terms. Disease expert reviewers will be asked to consider the suitability of the suggested terms and suggest other core terms that should be included. These terms will be tagged as core terms in the Genomics England catalogue.

3. Aim to avoid very common terms or very rare terms

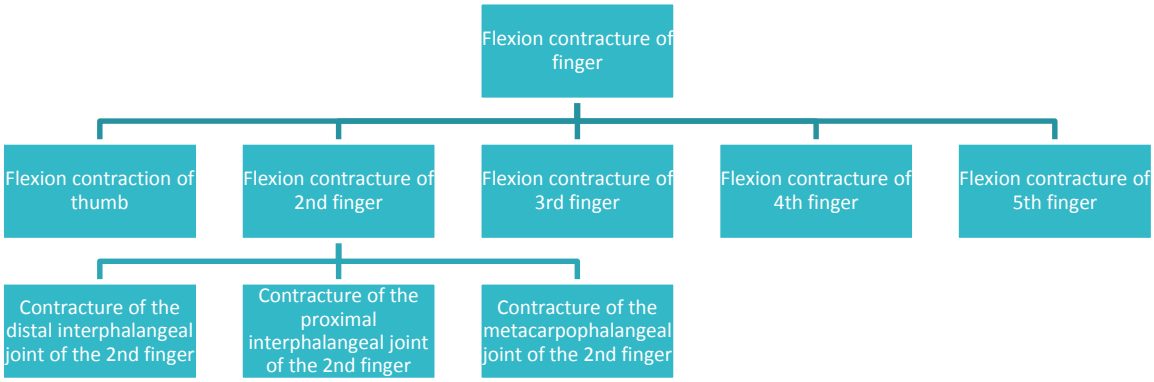
More general terms – i.e. those which are near to the top level of HPO's hierarchy, such as 'Abnormality of the skeletal system' - do not add much information to a disease model and may capture misleading data due to the large number of descendent terms that they represent. For example, stating that "Abnormality of the skeletal system" is absent in a participant implies that all descendent HPO terms with any link to the skeletal system have been assessed and also declared absent. More detailed terms – i.e. those lower down the HPO hierarchy – are more informative, and should be favoured over very high level terms if possible.

In contrast, in some cases a phenotype has been associated with a disease but is rarely if ever seen in practice. If it is unlikely to ever be recorded it will not add any information to the disease model and should be considered for removal. Recruiters are always able to search and include any additional HPO terms which they see in a participant so it will still be possible to capture this phenotype. Of course we recognise that the 100,000 genome programme addresses diseases where extremely rare phenotypes may be highly relevant, so the emphasis will always be to retain detailed HPO terms and avoid general terms where possible.

Achieving this: where models are too large, or when the Monarch phenotype sufficiency score (see above) is low, the Genomics England team will manually review the terms with the disease expert teams to identify those which are very common, and possibly those which are extremely rare, as candidates for exclusion. When developing models it should always be the aim to choose phenotypes that are detailed rather than general.

4. Avoid redundancy or unnecessary duplication

HPO is an extensive and hierarchical terminology. An abnormality such as 'Flexion contracture of finger' is broken down into joint contractures of the first, second finger etc., and for each of these into contractures of each finger joint:



Models can be designed with terms at any of these levels of detail, but including all the terms at all levels is likely to be excessive and may not make the model more informative. In some cases it risks collecting data that are contradictory.

However an important consideration is that recruiting clinicians may need to record phenotype data at different levels of detail for different participants, so will need to have HPO terms at more than one level of detail.

Achieving this: the Genomics England team will analyse larger models to show where a single higher-level HPO term can be added to replace two or more related lower-level terms, or where related higher and lower-level terms are present in a model. These will be fed back to disease expert teams to determine if they represent genuine redundancies and if modifications can be made without losing detail in the model.

5. Aim for consistency of approach between disease models

In order to improve comparability of data between diseases, we aim for consistency of approach to phenotypes present in different disease models. This is particularly relevant for phenotypes that can be captured in multiple different ways in the ontology or are overlapping, for example developmental delay and intellectual disability.

Achieving this: the Genomics England team will highlight phenotypes present in multiple models and suggest a shared core approach. These will be presented to the disease expert teams to guide model revision.