



King's Research Portal

Document Version

Publisher's PDF, also known as Version of record

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Kudama, S., Berlanga, R., Houlden, H., Jiménez-Ruiz, E., Jonvik, H., Milward, A., Morris, H., Ryten, M., Saklatvala, J., Simpson, M. A., & Wood, N. (2015). Towards enabling the semantic access of phenotypic information in clinical letters. *CEUR Workshop Proceedings*, 1546, 177-178.

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Towards Enabling the Semantic Access of Phenotypic Information in Clinical Letters

Shahad Kudama¹, Rafael Berlanga¹, Henry Houlden² Ernesto Jiménez-Ruiz³, Hallgeir Jonvik², Adam Milward⁴, Huw Morris², Mina Ryten⁴, Jake Saklatvala⁵, Michael A. Simpson⁵, and Nicholas Wood²

¹Universitat Jaume I, Spain ²University College London, UK ³University of Oxford, UK
⁴Genomics England, UK ⁵King's College London, UK

1 Motivation

Over the past 3 years there has been a massive growth in genetic testing both in terms of the scope of testing and the numbers of individuals offered genetic testing. Targeted sequencing of small genomic regions has been replaced by panel testing, whole exome sequencing and most recently whole genome sequencing [4]. Furthermore, genetic testing in research, but also clinical settings has extended beyond small numbers of selected individuals with very rare, highly-defined disorders to cover larger populations.

While this information presents new clinical opportunities and opens the way for development of novel therapies, it also presents major challenges. For clinicians, reliable identification of disease-associated genetic variants from amongst the broader background of variants present in all human genomes that are rare, but not actually pathogenic, is a concern. It is likely that for many rare genetic disorders obtaining clarity will require a worldwide effort and so the ability to capture and share key clinical information, as well as genetic information, is becoming increasingly important.

However, at present there are major challenges with regard to the collection and storage of clinical information, particularly in the context of rare genetic disorders. The process of studying a patient with a possible rare genetic disorder typically involves many different clinical specialists with no “standard” patient route. During this process, it is very common to refer from one specialist to another in order to obtain a range of opinions and access different tests. The output of this process is usually clinical letters which are used both to document the patient’s progress and communicate findings between specialists (e.g. patient history, examination findings, investigation results and clinical impression).

Since clinical letters are a key source of knowledge, their proper annotation and storage would enable access to the information they contain in a systematic way.

Currently, the creation and processing of clinical letters requires the following steps:

1. Letters are dictated using a speech recognition system by the clinician.
2. The recording is uploaded to a server.
3. The voice data is transcribed using another application.
4. The text is downloaded and checked by qualified personnel.
5. The letter is tagged manually by a specialist responsible for reading and annotating the interesting terms.

This final part of the process requires input from a qualified medical professional and is often time-consuming. This has led to the relatively limited use of clinical letters and consequently the loss of potentially important information. Thus, suitable software support is required to assist the clinician in the process of annotating and processing clinic letters to make this a part of “normal” working practice.

The screenshot displays the PHENOTAG main interface. On the left, a 'Color legend' lists categories: Persons and personal relations (blue), Observations (red), HPO (purple), Shifters (green), Modifiers (yellow), and Genes/Anatomy (orange). Below this is the 'HPO Information Content Score: 5 / 424 = 1.18 %'. The main area shows a snippet of a clinical letter with HPO terms highlighted in color. On the right, a list of HPO terms for patient 'K' is shown, including 'carry ||| 15q11.2', 'deletion ||| 15q11.2', 'deletions ||| 15q11.2', 'epilepsy || risk |', 'learning difficulties |||', 'difficulties and behavioural | difficulties |', 'behavioural problems |||', 'carries ||| PAR1', 'duplication ||| PAR1', 'epilepsy |||', and 'duplication || identified | PAR1'. Other patients 'B' and 'E' are also listed with their respective HPO terms. A 'Download data' button is at the bottom right.

Fig. 1. PHENOTAG main interface

2 Proposed solution

In this paper we introduce PHENOTAG, a prototype software aimed at allowing the annotation and storage of phenotypic information from clinical letters with the aim of improving the accuracy of genetic testing. PHENOTAG has the added innovation of allowing specialists to visualise the information content of their letters and potentially check the quality of annotation within their normal workflow. Figure 1 shows an overview of PHENOTAG's interface. The annotator underlying PHENOTAG is currently based on the approaches presented in [2, 5] and it can be used in conjunction with a series of coordinated vocabularies (e.g., [3]) making special emphasis in HPO concepts [1].

We envisage the following uses for this software:

1. The annotation and storage of HPO terms contained within clinic letters for the purposes of populating predefined disease-specific data models for a wide range of diseases (of the kind being developed by the 100,000 Genomes project).
2. The annotation and storage of HPO/other pre-defined terms contained within clinic letters for patients consented for research into the genetics of a single disorder (e.g. Parkinson's Disease).
3. The annotation and storage of HPO/other predefined terms contained within clinic letters in order to assess the information content and quality of clinical documentation within a clinical genetics service.

References

1. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 42(D1), D966–D974 (Jan 2014)
2. Berlanga, R., Nebot, V., Jiménez, E.: Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural* 45, 247–250 (2010)
3. Jiménez-Ruiz, E., Grau, B.C., Llavori, R.B., Rebholz-Schuhmann, D.: First steps in the logic-based assessment of post-composed phenotypic descriptions. In: *SWAT4LS* (2010)
4. Katsanis, S.H., Katsanis, N.: Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet* 14(6), 415–426 (Jun 2013)
5. Nebot, V., Berlanga, R.: Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. *Knowl. Inf. Syst.* 38(2), 365–389 (2014)