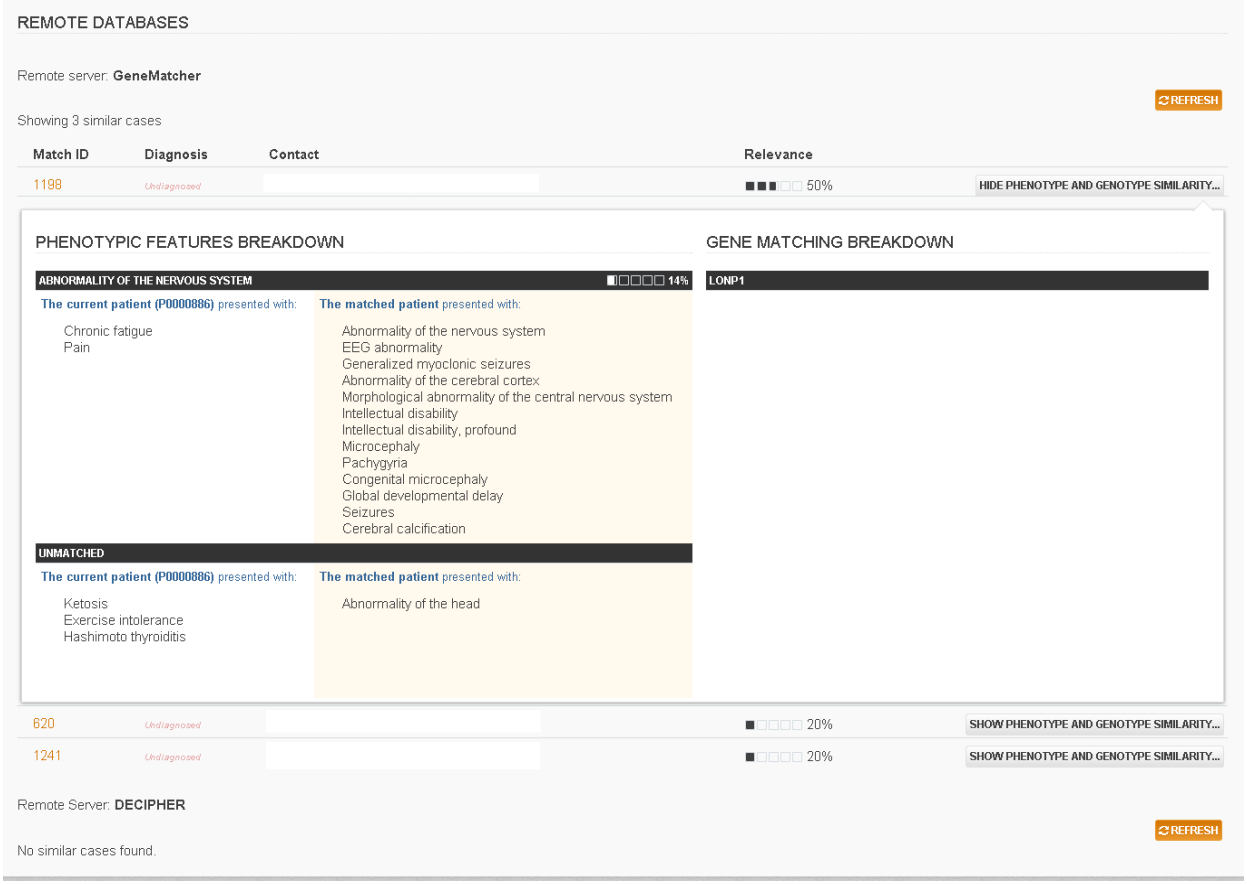
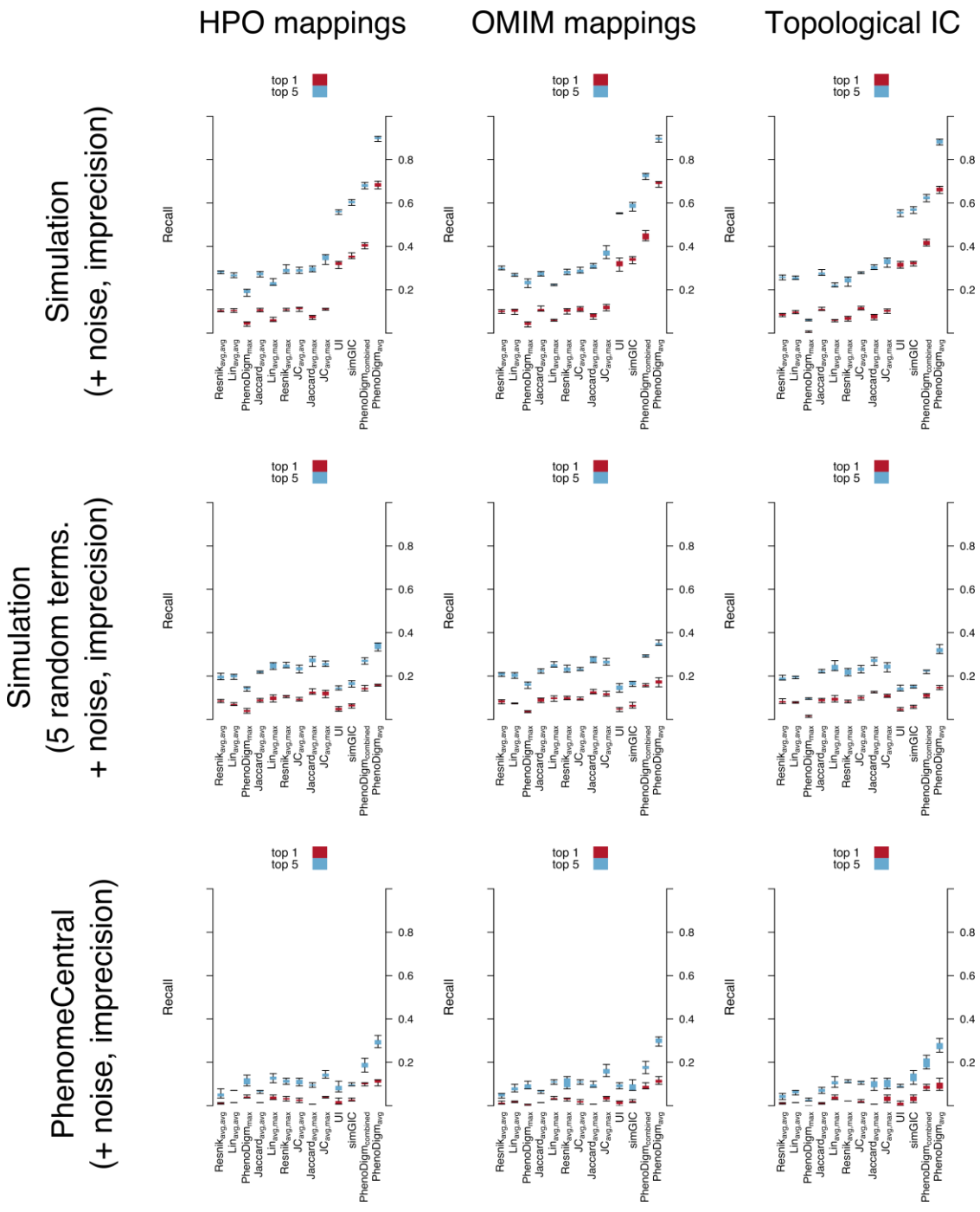


Supp. Figure S1. The user interface in PhenomeCentral for showing similar patients in remote databases using the Matchmaker Exchange API. Submitter details have been redacted.



Supp. Figure S2. The effect of different methods for information content calculation on the performance of each phenotypic similarity measure on simulated patients with noise added (top and middle rows) and real patients with noise added (bottom row). The information content was calculated in three ways: based on the disease-phenotype mappings provided by the HPO (left column), based on the disease-phenotype mappings provided by OMIM (center column), and based only on the topology of the HPO as using the same method as GeneYenta (right column). The overall performance of most measures appears to be robust to these differences.

Supp. Table S1. Comparison of the performance of 13 similarity measures in their ability to find patients with the same rare disease based on the HPO terms annotated for each patient

Measure	Versions	Definition	Reference
Resnik	avg, max	$Resnik(a,b) = max_{t \in g^a \cap g^b} IC(t)$	see the review by (Pesquita et al., 2009)
Lin	avg, max	$Lin(a,b) = \frac{2 * Resnik(a,b)}{IC(a) + IC(b)}$	
JC	avg, max	$JC(a,b) = \frac{1}{IC(a) + IC(b) - 2 * Resnik(a,b) + 1}$	
Jaccard	avg, max	$Jaccard(a,b) = \frac{\lvert g^a \cap g^b \rvert}{\lvert g^a \cup g^b \rvert}$	
UI		$UI(P,Q) = \frac{\lvert g^P \cap g^Q \rvert}{\lvert g^P \cup g^Q \rvert}$	
PhenoDigm	avg, max, combined	$PhenoDigm(P,Q) =$ see reference	(Smedley et al., 2013)
simGIC		$sim_{GIC}(P,Q) = \frac{\sum_{t \in g^P \cap g^Q} IC(t)}{\sum_{t \in g^P \cup g^Q} IC(t)}$	(Pesquita et al., 2007)

The *Resnik*, *JC*, *Lin*, and *Jaccard* measures compare two ontology terms, a and b. To measure the similarity between two patients (i.e. between two sets of ontology terms, P and Q), either the average score (avg) or best score (max) for each term in P is averaged together. The smoothed reciprocal of the *JC* distance measure was used as a similarity measure. In contrast, the *UI*, *PhenoDigm*, and *simGIC* measures directly score two sets on ontology terms. Three variants of the *PhenoDigm* score are described in (Smedley et al., 2013), and all three were included in the evaluation. The information content of a term is defined as $IC(t) = log(p(t))$ where $p(t)$ is the fraction of all disease-HPO mappings that involve term t (or a descendant of t). We also compared this to a topological definition: $IC(t) = (|g_t| + 1)/N$, where N is the number of terms in the HPO and g_t is the set of terms including t and all descendants of t. In the table, g^t is the set of terms induced by t (the set of nodes including t and all ancestors of t), and g^P is the set of terms induced by the set of terms in patient P.