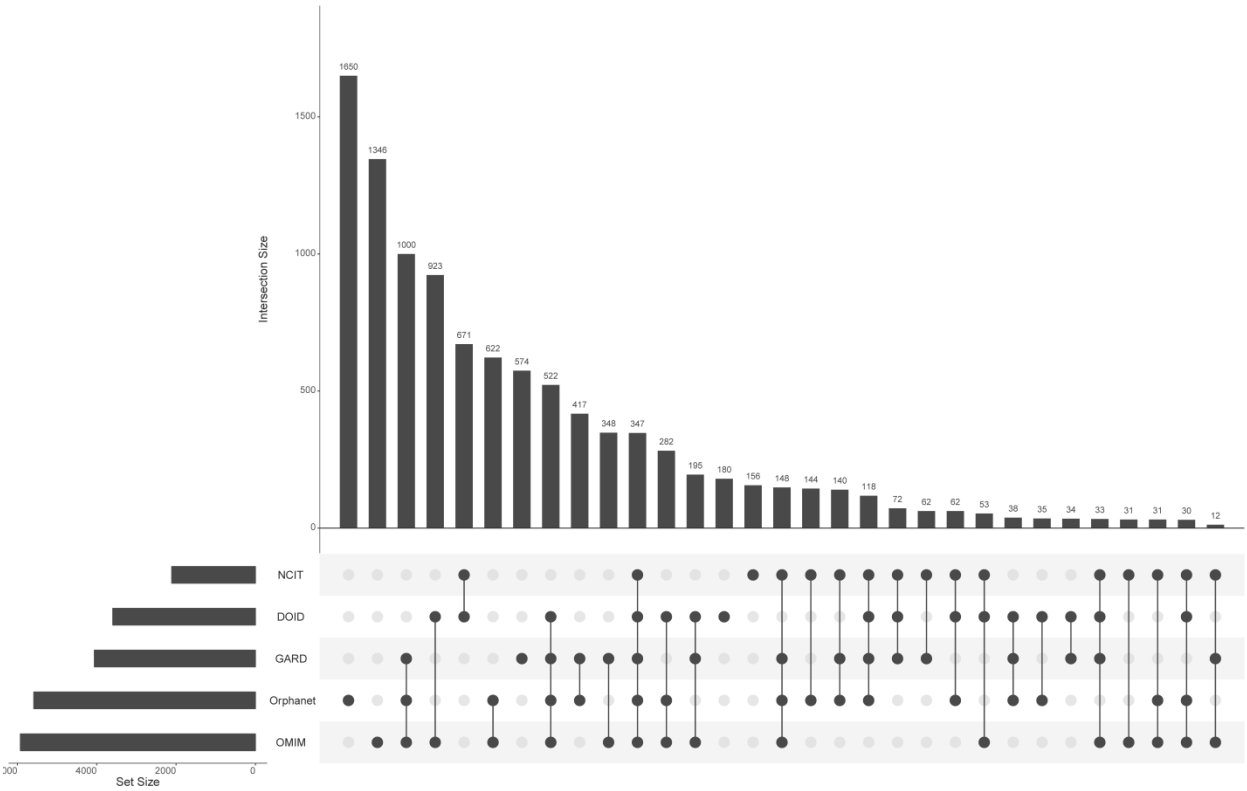
Supplementary Box 1 | **Definitions, data curation and analysis to estimate the number of rare diseases**

Disease concepts and leaf terms
In nosology, ‘disease concepts’ are concepts that define a disease by its specifically associated symptoms and pathognomonic signs and, more recently, by the underlying cause such as a gene mutation or other causative agents, such as viruses. Disease-centric terminological resources such as the Monarch Disease Ontology (Mondo) are currently focused on harmonizing representations of diseases. The ontology tree in such resources is comprised of nodes at various levels of hierarchy. Terminal nodes — that is, the most specific terms (‘disease concepts’) that have no lower-level (‘child’) terms below — are referred to as ‘leaf terms’.

In Mondo, each distinct disease is represented by a single identifier denoting the conceptualization of that disease. Whenever disease identifiers from different resources (for example, OMIM and OrphaNet) are found to be equivalent based upon a number of attributes, they are merged into the same disease concept. If they are broader or narrower, they are categorized as separate but related concepts.

Curation process for counting rare diseases
By counting ‘leaf terms’ that are also ‘rare diseases’ in Mondo, and thus excluding interior nodes (higher-level terms), we estimated the actual number of rare diseases. The procedure is described in detail at http://doi.org/10.5281/zenodo.3478576 and the output is shown in the figure below:



Supplementary Figure | Overlap and unique rare diseases concepts recorded in different knowledge sources. The horizontal bars on the left represent the total number of disease concepts in each of five disease resources. The vertical bars count disease “leaf” concepts and are sorted in descending order. The dots highlight the visual overlap between the five sources within the harmonized Mondo resource. OrphaNet (1,650) and OMIM (1,346) have the highest number of “unique” rare disease leaf terms, with the next highest showing overlap between GARD, Orphanet and OMIM for 1,000 unique rare disease leaf terms. Only 347 rare diseases match across all five resources (5 black dots). We consider that the combination of all five resources provides a much better approximation of the number of rare diseases than historical values. Some questions and caveats are discussed below.

Is it possible that different leaf terms in different databases actually describe the same disease?
The algorithmic process specifically designed for Mondo is used to suggest to the human curators that certain disease concepts should be merged. It intentionally errs on the side of conservatism, so there is a chance that some disease concepts should be merged. Our current estimate does take into account a slow-down in the rate of concept merges performed by human curators, which is indicative of an asymptotic point. It should be emphasized that this is work in progress, and that more merges, splits and combinations of the two are likely to happen in the future. The phenomenon of "lumping and splitting" in nosology has been described in 1969[1].

Criteria used to determine that a disease is rare
Given the current levels of annotation, it is not possible to authenticate that all diseases counted here are rare. Some of the resources incorporated in this effort do include non-rare diseases. Our procedure for classifying a disease as rare is inclusive, namely we keep the attribute 'rare' as provided by any of the contributing resources. For example, a disease is considered rare by default if catalogued in OrphaNet, despite the fact that the European definition of 'rare' differs from the US definition. We further consider any Mendelian disease (for example, indexed in OMIM) caused by variants in a single gene as 'rare'. This procedure is intentionally designed for exploring the entire set of rare diseases. Some of the definition criteria may be changed in the future, since they will depend on contextual definitions of rareness. The main rationale for this article is that no single definition of what constitutes a rare disease exists. Here we outline the technical framework to apply a single definition across all sources.

Inclusion of rare cancers
Cancer-predisposing conditions such as Lynch syndrome are typically included when counting rare diseases. A number of rare cancers are also included in this count, since the NCI Thesaurus is one of the sources for Mondo.

References
1. McKusick, V. A. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.* **12**, 298–312 (1969).