Supplemental table 2. Analysis of the Causes of False Negatives in MedGene

No. of genes	Incidence	The nature of false negatives
2	0.7%	Lack of MeSH Index
3	1.1%	Gene name only appeared in paper body and not in the abstract or title
6	2.1%	Gene name syntax was non-standard, e.g. “cystatin A,B and C”; “cystatin E/M”
12	4.2%	Gene symbols which were eliminated by specificity filters
2	0.7%	Gene term too new for current version
1	0.4%	Gene found only in review paper

Analysis of the causes of false negatives in MedGene. We compared the breast cancer-related genes at MedGene (2359 genes) with genes listed in other databases (285 genes) and identified 26 genes that were absent from MedGene. All of the available literature on these 26 genes was examined manually (about 80 papers) in order to determine the reason that the gene-disease association was missed by automated text mining. Each source of error and the number of genes that caused that error rate are indicated above.