Appendix 2: Cosine Similarity Calculations

The anaphylaxis query vector $\vec{V}(q)$ would consist of all the medical terms in the major and minor criteria of the BC case definition as well as the anaphylaxis term, while the report vector $\vec{V}(r)$ would consist of the corresponding extracted medical terms. Thus, we created a query vector of 33 components:

$\vec{V}(q) =$ [‘anaphylaxis’ ‘angioedema, localized or generalized’ ‘decreased level of consciousness’ ‘persistent, dry cough ‘cyanosis’ ‘diarrhoea’ ‘difficulty breathing’ ‘erythema, generalized’ ‘eyes, red and itchy’ ‘grunting’ ‘hypotension’ ‘nausea’ ‘pain, abdominal’ ‘prickle sensation’ ‘pruritus, generalized’ ‘respiratory distress’ ‘rhinorrhoea’ ‘skin rash’ ‘sneezing’ ‘stridor’ ‘swelling, upper airways’ ‘tachycardia’ ‘tachypnoea’ ‘throat closure, sensation of’ ‘uncompensated shock’ ‘urticaria (hives), generalized’ ‘voice, hoarse’ ‘vomiting, unspecified’ ‘wheezing(bronchospasm), bilateral’ ‘accessory respiratory muscles, increased use of’ ‘pulse volume, reduced central’ ‘urticaria (hives), localized at injection site’ ‘recessions’]

An example of a report vector $\vec{V}(r_1)$ with 6 anaphylaxis-related components would be:

$\vec{V}(r_1) =$ [‘erythema, generalized’ ‘eyes, red and itchy’ ‘pruritus, generalized’ ‘sneezing’ ‘throat closure, sensation of’ ‘urticaria (hives), generalized’]

IR suggests the formulation of numerical vectors through the assignment of certain weights to the vector components. To formulate the numerical query vector we calculated the weight of each term that was equal to the inverse report frequency (irf_t) of each term in the training set or equal to zero when no reports included this term. The irf_t (or the weight of each term in the query – $w_{t,q}$) was equal to:

$$w_{t,q} = log\frac{N}{rf_t}$$

where N was the number of reports in the training set (N=4526) and rf_t the number of reports that contained the term 't'. Subsequently the $w_{t,q}$ was length-normalized divided by the Euclidean length of the vector, e.g. for sneezing the $normalized\_w_{t_{sneezing},q}$ was:

$$normalized\_w_{t_{sneezing},q} = \frac{w_{t_{sneezing},q}}{\sqrt{(w_{t_1,q})^2 + (w_{t_2,q})^2 + \ldots + (w_{t_{33},q})^2}}$$

where t_1, t_2 and t_{33} were equal to 'anaphylaxis', 'angioedema, localized or generalized' and 'recessions', respectively.

For the report vectors, we used rf_t weighting with no use of irf_t but with Euclidean normalization. Thus the normalized weight of 'sneezing' in the above report $normalized\_w_{t_{sneezing},r_1}$ was:

$$normalized\_w_{t_{sneezing},r_1} = \frac{w_{t_{sneezing},r_1}}{\sqrt{(w_{t_1,r_1})^2 + \ldots + (w_{t_6,r_1})^2}}$$

where t1 and t6 were equal to 'erythema, generalized' and 'urticaria (hives), generalized', respectively.

The cosine similarity (or score) for the above report would be equal to:

$$sim(r_1, q) = \sum \left(normalized_{w_{t_1,q}} \times normalized_{w_{t_1,r_1}}\right) + \ldots + \left(normalized_{w_{t_{33},q}} \times normalized_{w_{t_6,r_1}}\right)$$

where t_8/t_1 and t_{26}/t_6 represented the terms ‘erythema, generalized’ and ‘urticaria (hives), generalized’. The subscripts denote the position of the components in the query and report vectors; obviously the same term is met in a different position per vector.

It should be also noted that the above equation is equivalent to the general equation used for the calculation of the cosine similarity of the report vs. the anaphylaxis query numerical vector:

$$sim(r_1, q) = \frac{\vec{V}(r_1) \cdot \vec{V}(q)}{\left|\vec{V}(r_1)\right| \cdot \left|\vec{V}(q)\right|}$$

where the numerator represents the dot product of the report and query vectors, while the denominator their Euclidean lengths.

The following table summarizes the score calculation for the above report.

BC-related terms in the report	**query**		**Report**		**product**
	rf_t	$w_{t,q}^*$	rf_t	$w_{t,r}^*$	$w_{t,q}$ x $w_{t,r}$
erythema, generalized	314	0.107	1	0.333	0.036
eyes, red and itchy	17	0.223	2	0.667	0.149
pruritus, generalized	455	0.092	1	0.333	0.031
Sneezing	32	0.198	1	0.333	0.066
throat closure, sensation of	210	0.123	1	0.333	0.041
urticaria (hives), generalized	426	0.094	1	0.333	0.031
Cosine similarity (score)					**0.353**

*normalized